ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ,ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΔΙΚΤΥΩΝ

Ανάπτυξη ενός ολοκληρωμένου διαδικτυακού περιβάλλοντος για την ανίχνευση δυνητικών λειτουργικών περιοχών γονιδίων

Διπλωματική εργασία

Θεοδωρίδη Στεφανία

Επιβλέποντες Καθηγητές :

Γεώργιος Σταμούλης

Νέστορας Ευμορφόπουλος

Ευχαριστίες

Αρχικά,θα ήθελα να ευχαριστήσω τον καθηγητή και επιβλέποντα της διπλωματικής μου εργασίας κ.Γεώργιο Σταμούλη για τις πολύτιμες συμβουλές και τη βοήθεια που μου παρείχε κατά τη διάρκεια των σπουδών μου και της εκπόνησης της διπλωματικής μου εργασίας.Επίσης να ευχαριστήσω τον κ.Ευμορφόπουλο για τη στήριξη και τη βοήθεια του κατά την περίοδο φοίτησης μου.Θα ήθελα τέλος να ευχαριστήσω την κ.Τριάδα Θηραίου ,η οποία με βοήθησε πολύ στην ολοκλήρωση της εργασίας μου.

Επιπλέον ένα μεγάλο ευχαριστώ στους φίλους μου για τις όμορφες στιγμές που περάσαμε κατά τη διάρκεια των σπουδών μας και για την υποστηρίξη τους κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.Τέλος θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένεια μου για αμέριστη συμπαράσταση, κατανόηση και υποστήριξη τους όλα αυτά τα χρόνια των σπουδών μου και κυρίως κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Περίληψη

Στόχος της συγκεκριμένης εφαρμογής είναι η σχεδίαση και η ανάπτυξη ενός ολοκληρωμένου διαδικτυακού περιβάλλοντος , στο οποίο ο κάθε χρήστης θα μπορεί να κάνει upload αρχεία που θα περιέχουν ακολουθίες DNA ,που ανήκουν σε δύο διαφορετικές ομάδες δεδομένων για επεξεργασία. Αρχικά, επιδιώκεται να γίνει μια γενική παρουσίαση βιολογικών εννοιών που αφορούν την Βιοπληροφορική. Επιπλέον, παρουσιάζονται αναλυτικά οι διαφορετικές μέθοδοι της που έχουν χρησιμοποιηθεί έως τώρα,όσον αφορά εργαλεία για εύρεση μοτίβων μέσα σε ακολουθίες γονιδίων.Στη συνέχεια παρουσίαζεται η εκτέλεση της εφαρμογής για δύο αρχεία ακολουθιών τα Up_seq_antagomir και τα other_seq_antagomir.

Πιο συγκεκριμένα στο 1^{O} Κεφάλαιο γίνεται μια γενικότερη περιγραφή των βασικών βιολογικών εννοιών που έχουν σχέση με το βιολογικό θέμα της εύρεσης μοτίβων και παρουσίαζονται ενδεικτικά κάποιες μέθοδοι.Στο 2^{o} Κεφάλαιο κάνουμε περιγραφή της εφαρμογής και του αλγορίθμου που χρησιμοποιείται ,περιγράφουμε τις μεθόδους και τα υλικά της εφαρμογής κάνοντας ανφορά στη γλώσσα Java ,Jsp και στο στατιστικό πακέτο που χρησιμοποιήθηκε.Στο 3^{o} Κεφάλαιο δίνουμε μια εικόνα χρήσης της εφαρμογής από τα διάφορα στάδια ,το διάγραμμα ροής της εφαρμογής καθώς και τα στατιστικά διαγράμματα.Τέλος στο 4^{o} Κεφάλαιο περιγράφονται τα συμπεράσματα μας από την υλοιποίηση της εφαρμογής σε σύγκριση με άλλες πειραματικές εφαρμογές και δίνονται ιδέες για επεκτάσεις της εφαρμογής.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Δεσοξυριβονουκλεϊκό οξύ(DNA)

Το δε(σ)οξυριβο(ζο)νουκλεϊ(νι)κό οξύ (Deoxyribonucleic acid - DNA) είναι ένα νουκλεϊκό οξύ που περιέχει τις γενετικές πληροφορίες που καθορίζουν τη βιολογική ανάπτυξη όλων των κυτταρικών μορφών ζωής_και των περισσοτέρων ιών.Το DNA συνήθως έχει τη μορφή διπλής έλικας.

Η αποκωδικοποίηση του DNA, η αποσαφήνιση δηλαδή του τρόπου με τον οποίο η δομή του DNA καθορίζει συγκεκριμένες γενετικές επιλογές, επέτρεψε στους επιστήμονες να κατανοήσουν καλύτερα την γενετική της ζωής και την κληρονόμηση ορισμένων χαρακτηριστικών και νόσων. Επειδή το DNA στα ορισμένα του σημεία είναι ξεχωριστό στον κάθε άνθρωπο, έχουν αναπτυχθεί μέθοδοι βασιζόμενες στην ταυτοποίηση του DNA και βρίσκουν εφαρμογή στην Ιατροδικαστική και στην Εγκληματολογία καθώς επίσης και στην αποσαφήνιση οικογενειακών σχέσεων μεταξύ ατόμων. Τα τελευταία χρόνια γίνεται η πιο εντατική η χρήση του DNA και στις μελέτες της ιστορίας και της ανθρωπολογίας.

Η ανακάλυψη της δομής του DNA πραγματοποιήθηκε το 1953 από τους Τζέιμς Γουάτσον (James D. Watson) και Φράνσις Κρικ (Francis Crick). Από πολλούς η ανακάλυψη της διπλής έλικας του DNA θεωρείται ως η μεγαλύτερη βιολογική ανακάλυψη του 20ου αιώνα. Για τη συνεισφορά τους στη μελέτη της δομής του DNA, οι Γουάτσον και Κρικ μοιράστηκαν το 1962 το Βραβείο Νόμπελ με τον Μόρις Γουίλκινς, ο οποίος εργάστηκε προς την ίδια κατεύθυνση.

Πρόκειται για μια μεγαλομοριακή ένωση που συγκροτείται από αζωτούχες-πρωτεϊνικές βάσεις, φωσφορικές ρίζες και ένα σάκχαρο με πέντε άτομα άνθρακα (πεντόζη), την δε(σ)οξυριβόζη. Στα ευκαρυωτικά κύτταρα ανιχνεύεται κυρίως μέσα στον πυρήνα του κυττάρου αλλά και σε μερικά άλλα οργανίδια, όπως τα μιτοχόνδρια και τα πλαστίδια, επιτρέποντάς τους να αναπαράγονται αυτόνομα (ημιαυτόνομα οργανίδια).

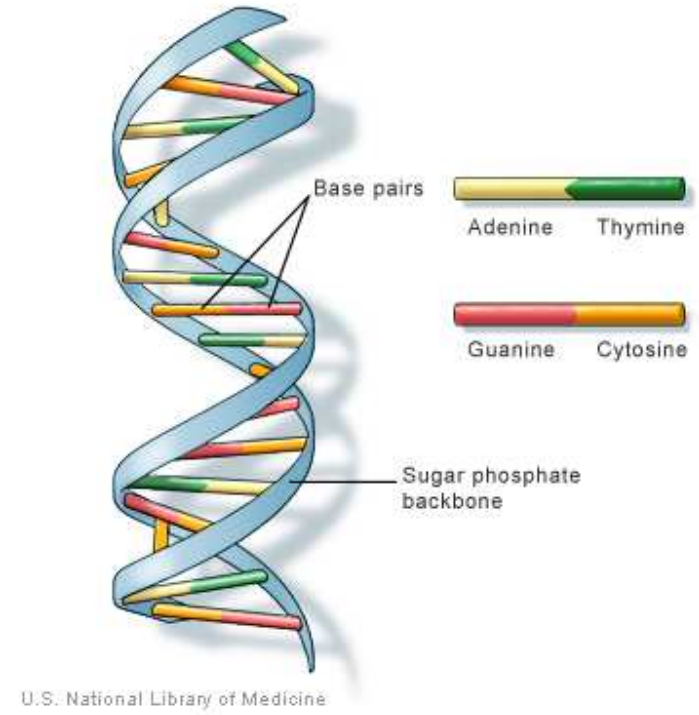
Το σύνολο των μορίων DNA που υπάρχουν σε ένα κύτταρο αποτελούν το γενετικό υλικό του. Το DNA είναι ο φορέας των γενετικών πληροφοριών του κυττάρου, όχι μόνον με την έννοια της μεταβίβασης χαρακτηριστικών, αναλοίωτων από γενεά σε γενεά, αλλά και της ρύθμισης της φυσιογνωμίας εξειδίκευσης κάθε κυττάρου για την επιτέλεση των ιδιαίτερων λειτουργιών του. Τέλος, το DNA επιτρέπει τη δημιουργία γενετικής ποικιλότητας, υφιστάμενο μεταλλάξεις.

- ***Δομή του DNA***

Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμήκων αλύσεων, οι οποίες συστρέφονται ελικοειδώς μεταξύ τους. Οι αζωτούχες βάσεις στο DNA είναι τέσσερις:

- κυτοσίνη C
- γουανίνη G
- θυμίνη T
- αδενίνη A

Οι αζωτούχες βάσεις, ανάλογα με την σειρά αλληλουχίας τους σε τριάδες, κωδικοποιούν το μήνυμα για τη σύνθεση των αμινοξέων του κυττάρου στα ριβοσώματα. Εκεί τα αμινοξέα συνδυάζονται, με τη σειρά κατά την οποία μεταφέρθηκαν στο ριβόσωμα και συντίθενται έτσι οι διαφορετικές πρωτεΐνες.



Εικόνα: Η δομή της διπλής έλικας του DNA

1.2 Ακολουθίες DNA

Η ακολουθία ή η πρωτογενής δομή ενός νουκλεϊκού οξέος , είναι ο ακριβής ορισμός της ατομικής του σύνθεσης και των χημικών δεσμών, που συνδέουν αυτά τα άτομα.Αφού τα νουκλεϊκά οξέα είναι μη διακλαδιζόμενα πολυμερή, αυτό ισοδυναμεί με τον ακριβή ορισμό της ακολουθίας των νουκλεοτιδίων που απαρτίζουν ολόκληρο το μόριο.Κατά συνθήκη η πρωτογενής δομή ενός μορίου DNA αναφέρεται από την περιοχή-άκρο 5’ ως την περιοχή-τέλος 3’.

Η ακολουθία έχει την ικανότητα να μεταφέρει πληροφορίες.Όταν χρησιμοποιείται για να αναφερθεί στο βιολογικό DNA ,το οποίο μεταφέρει πληροφορίες για να κατευθύνει τις λειτουργίες των έμβιων όντων, ο όρος γενετική ακολουθία χρησιμοποιείται συχνά.Οι ακολουθίες μπορούν να διαβαστούν μέσω των μεθόδων αλληλουχίας DNA.

Ο όρος ακολουθία DNA αναφέρεται σε μεθόδους αλληλουχίας για τον προσδιορισμό της σειράς των νουκλεοτιδικών βάσεων-αδενίνη, γουανίνη, κυτοσίνη,θυμίνη- στο μόριο ενός DNA.

Η γνώση των ακολουθιών DNA έχει γίνει απαραίτητη για τη βιολογική έρευνα , σε άλλους κλάδους της έρευνας που χρησιμοποιούν τις ακολουθίες DNA και σε πολυάριθμα εφαρμοσμένα πεδία όπως η διάγνωση, η βιοτεχνολογία,η εγκληματολογική βιολογία και η βιολογική συστηματική.Η αλληλουχία του DNA έχει επιταχύνει σημαντικά την βιολογική έρευνα και ανακάλυψη.Η ραγδαία πρόοδος του να παριστάται το DNA ως ακολουθία που έχει επιτευχθεί με τη σύγχρονη τεχνολογία της παράστασης του DNA ως ακολουθία, έχει συμβάλει σημαντικά σε μεθόδους για την ακολουθία του ανθρώπινου γονιδιώματος , στο Πρόγραμμα Ανθρωπίνου Γονιδιώματος. Σχετικές εργασίες συχνά με τη συνεργασία επιστημόνων από όλο τον κόσμο, έχουν δημιουργήσει τις ολοκληρωμένες ακολουθίες DNA πολλών ζώων ,φυτών και μικροβιακών γονιδιωμάτων.

Οι πρώτες ακολουθίες DNA αποκτήθηκαν στις αρχές της δεκαετίας του 70’ από ερευνητές της ακαδημαϊκής κοινότητας ,χρησιμοποιώντας επίπονες μεθόδους βασισμένες στη δισδιάστατη χρωματογραφία. Αργότερα η ανάπτυξη μεθόδων ακολουθίας με αυτοματοποιημένη ανάλυση έκανε τον προσδιορισμό της ακολουθίας του DNA πολύ πιο εύκολο και γρήγορο.

Μια ακολουθία DNA κωδικοποιιεί τις απαραίτητες πληροφορίες για την επιβίωση και την αναπαραγωγή των έμβιων όντων.Γι’αυτό τον λόγο ο προσδιοριμός της ακόλουθίας ,είναι χρήσιμος στην έρευνα στο γιατί και πώ ζουν οι οργανισμοί.Λόγω της στρατηγικής σημασίας του DNA στα έμβια όντα ,η γνώση της ακολουθία του DNA , μπορεί να είναι χρήσιμη σε σχεδόν οποιαδήποτε βιολογική έρευνα.Για παράδειγμα στον τομέα της ιατρικής μπορεί να χρησιμοποιηθεί για την ταυτοποίηση, τη διάγνωση και τη μελλοντική ανάπτυξη θεραπειών για γενετικές ασθένεις.Ομοίως η έρευνα σε παθογόνους παράγοντες μπορεί να οδηγήσει σε θεραπείες για μεταδοτικές ασθένειες.

1.3 Γονίδια

Το γονίδιο είναι μια μονάδα κληρονομικότητας σε έναν ζωντανό οργανισμό . Είναι κανονικά ένα τμήμα του DNA που κωδικοποιεί για έναν τύπο πρωτεΐνης ή για μια αλυσία RNA που έχει μια λειτουργικότητα στον οργανισμό.Όλες οι πρωτεΐνες και οι λειτουργικές αλυσίδες RNA καθορίζονται από τα γονίδια.Όλα τα έμβια όντα εξαρτώνται από τα γονίδια .Τα γονίδια κρατάνε τις πληροφορίες για να δημιουργήσουν και να διατηρήσουν τα κύτταρα ενός οργανισμού και να περάσουν τα γενετικά γνωρίσματα στους απογόνους.Ένας μοντέρνος ορισμός για ένα γονίδιο είναι : «μια εντοπίσιμη περιοχή της γονιδιωματικής ακολουθίας , αντίστοιχα με μια μονάδα κληρονομικότητας ,η οποία συνδέεται με τις ρυθμιστικές περιοχές ,τις μεταγραφικές περιοχές και άλλες λειτουργικές περιοχές της ακολουθίας».[1][2]

Η έννοια του γονιδίου ,εξελίσσεται με την επιστήμη της γενετικής , οποία ξεκίνησε όταν ο Gregor Mendel που παρατήρησε ότι οι βιολογικές μεταβολές κληρονομούνται από τους γονεϊκούς οργανισμούς ως ειδικά ,διακριτά γνωρίσματα.Η βιολογική οντότητα που είναι υπεύθυνη για τον ορισμό των γνωρισμάτων ,ονομάστηκε αργότερα γονίδιο,αλλά η βιολογική βάση για την κληρονομικότητα παρέμεινε άγνωστη μέχρι την αναγνώριση του DNA ως το γενετικό υλικό στη δεκαετία του 40'.Όλοι οι οργανισμοί έχουν πολλά γονίδια που αντιστοιχούν σε πολλά διαφορετικά γνωρίσματα ,μερικά από τα οποία είναι άμεσα ορατά όπως το χρώμα των ματιών,και άλλα όχι ,όπως ο τύπος αίματος ή ο αυξημένος κίνδυνος για συγκεκριμένες ασθένεις ή οι χιλιάδες βασικές βιοχημικές διαδικασίες που συντελούνται.

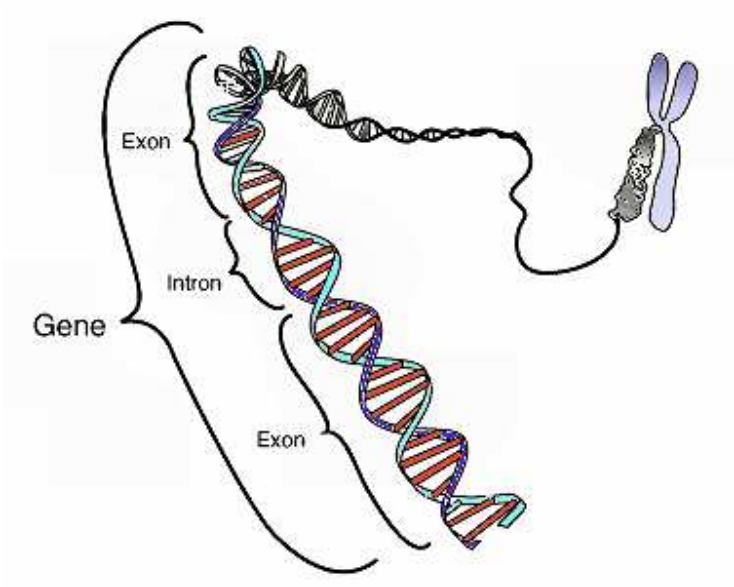
Η μεγάλη πλειονότητα των ζωντανών οργανισμών κωδικοποιούν τα γονίδιά τους στις μακριές έλικες του DNA.Η έκφραση των γονιδίων που κωδικοποιούνται στο DNA ,ξεκινάει από τη μεταγραφή του γονιδίου σε RNA ,έναν δεύτερο τύπο νουκλεϊκού οξέος που μοιάζει πολύ με το DNA ,αλλά του οποίου τα μονομερή περιέχουν τη ριβόζη αντί της δεσοκυριβόζης.Το RNA επίσης περιέχει τη βάση ουρακίλη στη θέση της θυμίνης.Τα μόρια RNA είναι λιγότερο σταθερά από αυτά του DNA και είναι μονόκλωνα.

Τα γονίδια που κωδικοποιούν πρωτεΐνες από μια σειρά τρι-νουκλεοτιδικών ακολουθιών που ονομάζονται κωδικόνια ,τα οποία χρησιμεύουν ως οι λέξεις στη γενετική γλώσσα.Ο γενετικός κώδικας καθορίζει την αλληλεπίδραση κατά τη μετάφραση της πρωτεΐνης μεταξύ των κωδικονίων και των αμινοξέων.Ο γενετικός κώδικας είναι περίπου ίδιος για όλους τους γνωστούς οργανισμούς.

1.3.1 Η λειτουργική δομή των γονιδίων

Όλα τα γονίδια έχουν ρυθμιστικές περιοχές πέρα από τις περιοχές που ρητά κωδικοποιούν για μια πρωτεΐνη ή ενα RNA προϊόν.Μια ρυθμιστική περιοχή που μοιράζεται σχεδόν σε όλα τα γονίδια είναι γνωστή ως ο promoter ,που προβλέπει μια θέση η οποία αναγνωρίζεται από τη μεταγραφική μηχανή ,όταν ένα γονίδιο πρόκειται να μεταγραφεί και να εκφραστεί. Ένα γονίδιο μπορεί να έχει περισσότερους από έναν promoters ,και έχει αποτέλεσμα σε RNAs που στο κατά πόσο μακριά εκτείνονται στο τέλος της περιοχής 5'.[5] Αν και οι περιοχές του υποκινητή έχουν μια ακολουθία συμφωνίας που είναι η πιο κοινή ακολουθία σε αυτή την περιοχή, μερικά γονίδια έχουν δυνατούς promoters που συνδέονται καλά στη μεταγραφική μηχανή και άλλους που έχουν «αδύναμους» promoters που δεν συνδέονται ισχυρά.Αυτοί οι αδύναμοι promoters συνήθως επιτρέπουν ένα χαμηλότερο ποσοστό της μεταγραφής από ότι οι ισχυροί προωθητές ,γιατί η μεταγραφική μηχανή συνδέεται μαζί τους και αρχίζει τη μεταγραφή λιγότερο συχνά.Άλλες δυνατές ρυθμιστικές περιοχές περιλαμβάνουν ενισχυτές ,οι οποίοι μπορούν να αντισταθμίσουν τους ασθενείς προωθητές.Οι περισσότερες ρυθμιστικές περιοχές είναι «upstream»,δηλαδή πριν ή προς το άκρο 5' του σημείου έναρξης της μεταγραφής.Οι ευκαρυωτικές προωθητικές περιοχές είναι πολύ περισσότερο δύσκολες να εντοπιστούν από ότι οι προκαρυωτικές.

Πολλά προκαρυωτικά γονίδια οργανώνονται σε οπερόνια, ή ομάδες γονιδίων των οποίων τα προϊόντα έχουν σχετικές λειτουργίες και τα οποία μεταγράφονται ως μονάδα. Αντίθετα τα ευκαρυωτικά γονίδια μεταγράφονται μόνο μια φορά ,αλλά μπορεί να περιλαμβάνουν μεγάλα τμήματα DNA που ονομάζονται ιντρόνια ,τα οποία μεταγράφονται αλλά ποτέ δεν μεταφράζονται σε πρωτεΐνη.



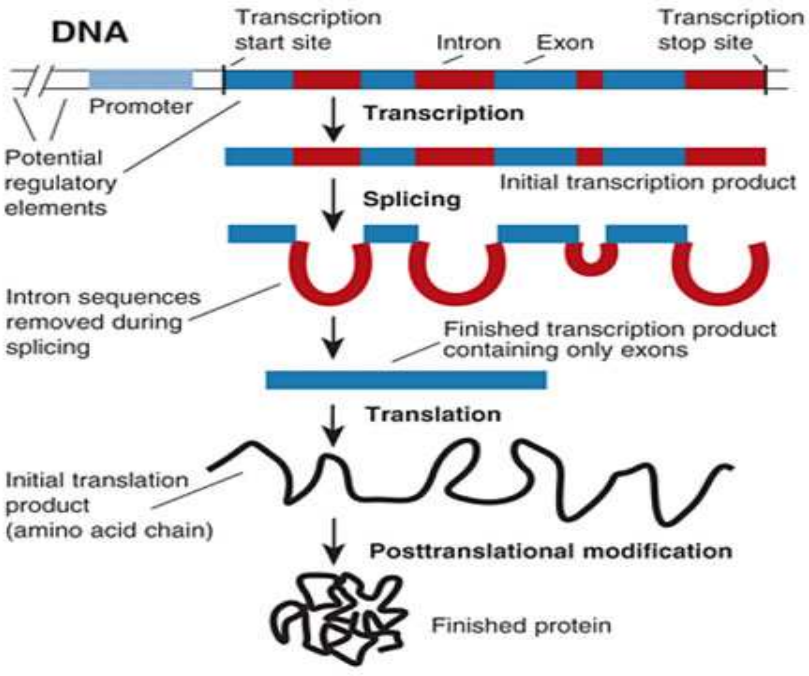
Εικόνα: το παραπάνω διάγραμμα δείχνει ένα γονίδιο σε σχέση με την δομή της διπλής έλικας του DNA και ενός χρωμοσώματος. Αυτό το διάγραμμα δείχνει μια περιοχή μόνο 50 περίπου βάσεων ως γονίδιο.Στην πραγματικότητα τα περισσότερα γονίδια είναι εκατό φορές μεγαλύτερα.

1.4 Γονιδιακή Έκφραση

Γονιδιακή έκφραση ονομάζεται η διαδικασία ,κατά την οποία οι πληροφρίες ενός γονιδίου χρησιμοποιούνται για τη σύνθεση ενός λειτουργικού γονιδιακού προϊόντος.Αυτά τα προϊόντα είναι συνήθως πρωτεΐνες ,αλλά τα γονίδια τα οποία δεν ειναι κωδικοποιητικά για πρωτείνες ,όπως το rRNA και το tRNA ,το παραγόμενο προϊόν τους είναι ένα λειτουργικό RNA. Η διαδικασία της γονιδιακής έκφρασης , χρησιμοποιείται από όλους σχεδόν τους οργανισμούς ,ευκαρυωτικούς (περιλαμβάνοντας του πολυ κυτταρικούς),προκαρυωτικούς και από τους ιούς ,για να δημιουργήσει τη μακρομοριακή μηχανή για τη ζωή.

Διάφορα βήματα στη διαδικασία της γονιδιακής έκφρασης μπορεί να διαφοροποιούνται, συμπεριλαμβάνοντας της μεταγραφή, το RNA splicing, τη μετάφραση και τη μετα-μεταγραφική τροποποίηση της πρωτεΐνης.Η γονιδιακή έκφραση δίνει τον έλεγχο στο κύτταρο πάνω στη δομή και τη λειτουργικότητα ,και είναι η βάση για την κυτταρική διαφοροποίηση ,την μορφογένεση ,την ευελιξία και την προσαρμοστικότητα κάθε οργανισμού.

Στη γενετική η γονιδιακή έκφραση ,είναι το πιο σημαντικό επίπεδο στο οποίο ο γενότυπος οδηγεί στον φαινότυπο. Ο γενετικός κώδικας μεταφράζεται από τη γονιδιακή έκφραση και οι ιδιότητες των προϊόντνων που παράγονται από την έκφραση ,δημιουργούν στον φαινότυπο του οργανισμού.



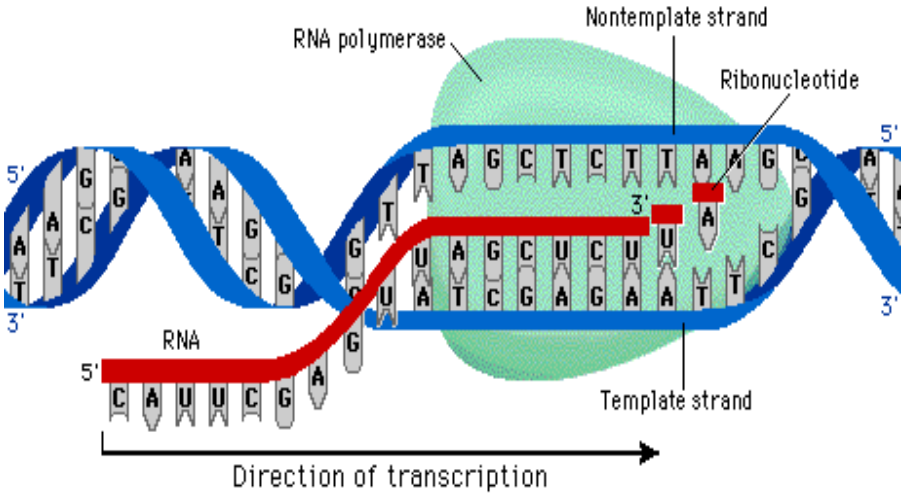
Εικόνα: γονιδιακή δομή και γονιδιακή έκφραση σε ανώτερους οργανισμούς

1.4.1 Μηχανισμός Γονιδιακής Έκφρασης

Μεταγραφή

Στη βιολογία, μεταγραφή ονομάζεται το πρώτο στάδιο της γονιδιακής έκφρασης και περιγράφει τη διαδικασία κατά την οποία δημιουργείται ένα μόριο RNA, με χρήση μιας αλυσίδας του DNA ως προτύπου, της οποίας είναι συμπληρωματικό. Χρησιμοποιείται ο όρος *μεταγραφή* γιατί η γενετική πληροφορία, στη γλώσσα του DNA, μεταγράφεται στη γλώσσα του RNA, με τη διαφορά πως η βάση ουρακίλη χρησιμοποιείται αντί της θυμίνης και βρίσκεται στη θέση όπου βρίσκεται η αδενίνη ,στην έλικα του DNA. (Το RNA συνίσταται από ουρακίλη στη θέση της θυμίνης). Η διαδικασία αυτή συμβαίνει στον πυρήνα των ευκαρυωτικών κυττάρων ή στο πυρηνοειδές των προκαρυωτικών. Σκοπός της είναι να μεταφερθούν οι γενετικές πληροφορίες από το DNA στα ριβοσώματα, για να γίνει η πρωτεϊνοσύνθεση. Η μεταγραφή του DNA μπορεί να γίνει πολλές φορές ταυτόχρονα επιταχύνοντας τις διεργασίες του κυττάρου. Σπανίως η μεταγραφή συμβαίνει αντίστροφα δημιουργώντας DNA με καλούπι το RNA από ρετροϊούς με τη βοήθεια του ενζύμου αντίστροφη μεταγραφάση.

Κατά τη μεταγραφή , το DNA διαβάζεται σπό την περιοχή 3'->5'. Εν τω μεταξύ το συμπληρωματικό RNA δημιουργείται προς την κατεύθυνση 5'->3'. Αν και το DNA είναι διατεταγμένο ως δύο αντιπαράλληλα σκέλη σε μια διπλή έλικα ,μόνο μία από τις έλικες του DNA ,που ονομάζεται πρότυπο σκέλος , χρησιμοποιείται για τη μεταγραφή.Αυτό συμβαίνει επειδή το RNA έχει μόνο μία έλικα ,σε ανίθεση με τη διπλή έλικα του DNA. Η άλλη έλικα του DNA καλείται κωδικοποιητκή έλικα , επειδή η ακολουθία της είναι ίδια με αυτή του RNA μετά τη μεταγραφή.(εκτός από την ύπαρξη της ουρακίλης στη θέση της θυμίνης)



Εικόνα:διαδικασία μεταγραφής

Επεξεργασία RNA

Η μεταγραφή των γονιδίων που κωδιοποιούν πρωτεΐνες ,δημιουργεί ένα αρχικό μεταγραφικό RNA στη θέση όπου βρισκόταν το γονίδιο.Αυτό το μεταγραφικό RNA μπορεί να μεταβληθεί πριν μεταφραστεί ,και είναι συγεκριμένα σύνηθες στους ευκαρυωτικούς οργανισμούς. Η πιο κοινή επξεργασία RNA είναι η αποκόλληση για να απομακρυνθούν τα ιντρόνια.

Η επεξεργασία RNA ,γνωστή και ως μετα-μεταγραφική τροποποίηση , μπορεί να αρχίσει κατά τη διάρκεια της μεταγραφής ,όπως στην περίπτωση της αποκόλλησης ,όπου απομακρύνονται τα ιντρόνια από το νέο κατασκευασμένο RNA.

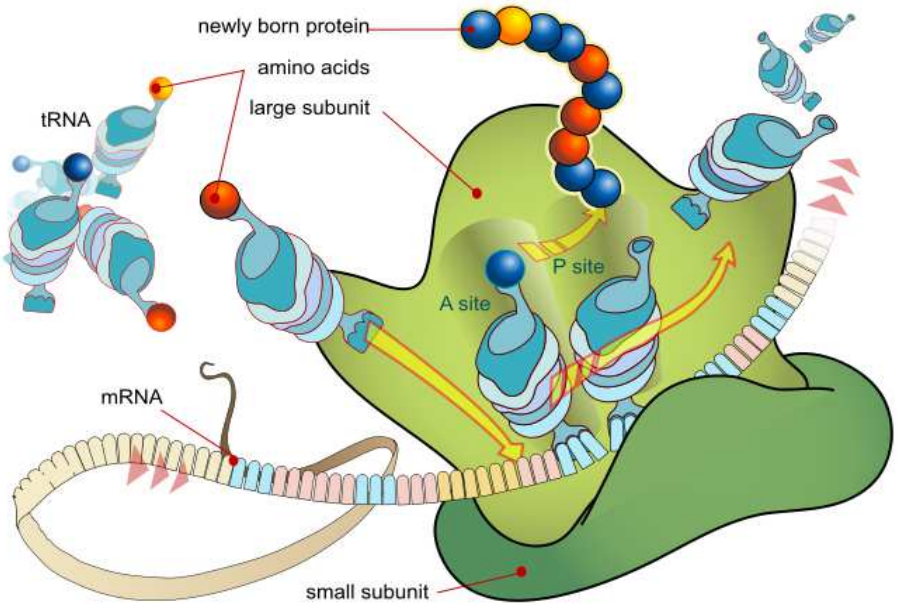
Η εκτεταμένη επεξεργασία του RNA μπορεί να είναι ένα εξελικτικό αποτέλεσμα ,που καθίσταται δυαντό από τον πυρήνα των ευκαρυωτικών κυττάρων.Στα προκαρυωτικά η μεταγραφή και η μετάφραση συμβαίνουν ταυτόχρονα ,ενώ στα ευκαρυωτικά η πυρηνική μεμβράνη χωρίζει τις δύο διαδικασίες δίνοντας χρόνο ,να πραγματοποιηθεί η επεξεργασία του RNA.

Μετάφραση

Για ορισμένα RNA (μη κωδικοποιητικά RNAs) το ώριμο RNA είναι το ολοκληρωμένο γονιδιακό προϊόν.Στην περίπτωση του αγγελιοφόρου RNA (mRNA) το RNA είναι ένας φορέας πληροφοριών για την κωδικοποίηση της σύνθεσης μίας ή περισσότερων πρωτεϊνών.

Κάθε τριπλέτα νουκλεοτιδίων στην κωδικοποιητική περιοχή του mRNA αντιστοιχεί σε ένα σημείο σύνδεσης για ένα μεταφορικό RNA. Τα μεταφορικά RNAs μεταφέρουν αμινοξέα , τα οποία δημιουργούν μια αλυσίδα στο ριβόσωμα.Το ριβόσωμα βοηθαέι το μεταφορικό RNA να συνδεθεί με το αγγελιοφόρο RNA και παίρνει το αμινοξύ από κάθε μεταφορικό RNA και φτιάχνει μια λιγότερο δομημένη πρωτεΐνη.

Στα προκαρυωτικά η μετάφραση συνήθως συμβαίνει μαζί με τη μεταγραφή ,χρησιμοποιώντας συχνά ένα αγγελιοφόρο RNA το οποίο βρίσκεται ακόμα στη φάση της δημιουργίας.Στα ευκαρυωτικά η μετάφραση μπορεί να συμβεί σε ένα εύρος περιοχών του κυττάρου ,ανάλογα με το που θα πρέπει να βρίσκεται η πρωτεΐνη που γράφεται. Σημαντικές θέσεις είναι το κυτταρόπλασμα και το ενδοπλασματικό δίκτυο. Αυτό διέπεται από το σωματίδιο αναγνώρισης σήματος –μια πρωτεΐνη που συνδέεται με το ριβόσωμακαι την κατευθύνει προς το ενδοπλασματικό δίκτυο, όταν βρίσκει μια ενδεικτική ακολουθία στην αυξανόμενη αλυσίδα των αμινοξέων.



Εικόνα:μετάφραση του mRNA από το ριβόσωμα

Μεταφορά πρωτεΐνης

Πολλές πρωτεΐνες προορίζονται για άλλα μέρη του κυττάρου εκτός από το κυτταρόπλασμα και μένα ευρύ φάσμα ενδεικτικών ακολουθιών χρησιμοποιούνται για να κατευθύνουν τις πρωτεΐνες εκεί που πρέπει να βρίσκονται.Στα προκαρυωτικά αυτό είναι μια απλή διαδικασία των περιορισμών του κυττάρου. Ωστόσο στους ευκαρυωτικούς ,υπάρχει μια μεγάλη ποικιλία διαδικασιών με διαφορετικούς ρόλους για να επιβεβαιώσουν πως η πρωτεΐνη θα φτάσει στο σωστό οργανίδιο.

Οι πρωτεΐνες δεν παραμένουν όλες μέσα στο κύτταρο και πολλές εξάγονται, για παράδειγμα πεπτικά ένζυμα,ορμόνες και εξωκυτταρικές πρωτεΐνες. Στα ευκαρυωτικά το μονοπάτι εξαγωγής είναι πολύ καλά ανεπτυγμένο και ο κύριος μηχανισμός της εξαγωγήςαυτών τνω πρωτεϊνών είναι μια μετατόπιση προς το ενδοπλασματικό δίκτυο.

Μέτρηση της γονιδιακής έκφρασης

Η μέτρηση της γονιδιακής έκφρασης είναι ένα πολύ σημαντικό κομμάτι σε πολλές επιστήμες που ασχολούνται με την υγεία και τη ζωή. Η δυνατότητα να υπολογιστεί το επίπεδο στο οποίο ένα συγκεκριμένο γονίδιο εκφράζεται μέσα στο κύτταρο ,στους ιστούς ή τον οργανισμό μπορεί να δώσει μια τεράστια ποσότητα πληροφορίας.Για παράδειγμα η μέτρηση της γονιδιακής έκφρασης μπορεί:

- Να εντοπίσει ιογενή λοίμωξη του κυττάρου (ιική έκφραση πρωτεϊνών)
- Να προσδιρίσει την ευαισθησία ενός ατόμου στον καρκίνο (έκφραση ογκογονιδίων)
- Να ανακαλύψει αν ένα βακτήριο είναι ανθεκτικό στην πενικιλίνη

Ομοίως η ανάλυση της θέσης της πρωτεϊνης που εκφράζεται αποτελεί ένα ισχυρό εργαλείο και αυτό μπορεί να γίνει σε έναν οργανισμό ή σε μια κυτταρική κλίμακα. Η διερεύνηση της τοποθεσίας είναι ιδιαίτερα σημαντική για τη μελέτη πολυκυτταρικών οργανισμών και ώς ένας δείκτης της λειτουργικότητας των πρωτεϊνών σε μονά κύτταρα.Ιδανικά η μέτρηση της γονιδιακής έκφρασης μπορεί να γίνει ανιχνεύοντας το τελικό προϊόν του γονιδίου, ωστόσο είναι συχνά πιο εύκολο να ανιχνευτεί ένας από τους προδρόμους ,συνήθως το mRNA και να ενδείξει το επίπεδο της γονιδιακής έκφρασης.

Σύστημα της γονιδιακής έκφρασης

Ένα σύστημα έκφρασης είναι ένα σύστημα ειδικά σχεδιασμένο ενός γονιδιακού προϊόντος κατ'επιλογή. Ένα σύστημα έκφρασης αποτελείται από ένα γονίδιο, συνήθως κωδικοποιημένο από το DNA ,και τη μοριακή μηχανή που απαιτείται για να μεταγραφεί το DNA σε mRNA και να μεταφράσει το mRNA σε πρωτεΐνη.Με την ευρεία έννοια αυτό περιλαμβάνει κάθε ζωνατνό κύτταρο αλλά ο όρος χρησιμοποιείται για να αναφερθεί ως ένα εργαστηριακό εργαλείο.Έτσι ένα σύστημα έκφρασης είναι τεχνητό κατά κάποιο τρόπο.Τα συστήματα έκφρασης είναι ωστόσο ,μια θεμελι΄ψδης φυσική διαδικασία.Οι ιοί είναι ένα άριστο παράδειγμα όπου αναπαράγονται κύτταρα ξενιστές ,ως ένα σύστημα έκφρασης για τις ιικές πρωτεΐνες και γονιδιώματα.

Τεχνικές Γονιδιακής έκφρασης

Οι ακόλουθες πειραματικές τεχνικές χρησιμοποιούνται για να μετρήσουν τη γονιδιακή έκφραση και είναι κατεταγμένες κατά χρονολογική σειρά,ξεκινώντας από την πιο παλιά,πιο καθιερωμένες τεχνολογίες.Χωρίζονται σε δύο ομάδες ,ανάλογα με το βαθμό της πολυπλοκότητας τους:

- Τεχνικές χαμηλής πολυπλοκότητας
    - Reporter γονιδίου
    - Northern Blot
    - Western Blot
    - Φθορισμός υβριδισμού
    - Αντίστροφη μεταγραφή

- Τεχνικές υψηλής πολυπλοκότητας
    - SAGE
    - Μικροσυστοιχίες DNA
    - Πίνακας πλακιδίων
    - RNA-seq

1.5 Ρύθμιση της γονιδιακής έκφρασης

1.5.1 Ρυθμιστικές ακολουθίες και ποιος ο ρόλος τους στη γονιδιακή έκφραση

Μια ρυθμιστική ακολουθία,(γνωστή και ως ρυθμιστική περιοχή) είναι ένα τμήμα του DNA ,όπου οι ρυθμιστικές πρωτεΐνες όπως οι μεταγραφικοί παράγοντες συνδέονται κατά προτίμηση.Οι ρυθμιστικές αυτές πρωτεΐνες συνδέονται σε μικρά τμήματα του DNA που ονομάζονται ρυθμιστικές περιοχές ,τα οποία είναι σωστά τοποθετημένα στο γονιδίωμα,συνήθως σε μικρή «upstream» απόσταση από το γονίδιο το οποίο ρυθμίζεται.Με αυτόν τον τρόπο αυτές οι ρυθμιστικές πρωτεΐνες μπορούν να προσλάβουν ένα άλλο συγκρότημα ,που ονομάζεται RNA πολυμεράση.Με αυτόν τον τρόπο ελέγχουν την γονιδιακή έκφραση και κατ'επέκταση την έκφραση της πρωτεΐνης.

Οι ρυθμιστικές ακολουθίες μπορούν επίσης να βρεθούν στο αγγελιοφόρο RNA, αλλά γενικά δεν έχουν μελετηθεί τόσο πολύ όπως αυτά του DNA. Ίσως να συνδέονται με πρωτεΐνες που συνδέονται με το RNA ή με άλλα RNAs.(miRNAs)

Η έρευνα για την έυρεση και άλλων ρυθμιστικών περιοχών στα γονιδιώματα σε όλες τις κατηγορίες οργανισμών ,βρίσκεται υπό εξέλιξη.

Η ρύθμιση της γονιδιακής έκφρασης αναφέρεται στον έλεγχο και στην ώρα εμφάνισης του λειτουργικού προϊόντος ενός γονιδίου. Ο έλεγχος της έκφρασης είναι ζωτικής σημασίας για να επτρέψει στο κύτταρο να παράγει τα γονιδιακά πρΐόντα που χρειάζεται όταν τα χρειάζεται.Με τη σειρά του αυτό δίνει στα κύτταρα την ευελιξία να προσαρμοστούν σε ένα μεταβλητό περιβάλλον ,σε εξωτερικά σημάδια ,σε βλάβη του κυττάρου,κλπ. Μερικά απλά παραδείγματα ,όπου η γονιδιακή έκφραση είναι σημαντικη είναι:

- Ο έλεγχος της έκφρασης της ινσουλίνης οπότε δίνει σήμα για τη ρύθμιση της γλυκόζης στο αίμα.
- Την αδρανοποίηση του χρωμοσώματος στα θηλυκά θηλαστικά για να αποτραπεί μια «υπερβολική δόση» των γονιδίων που περιέχει.

Κάθε βήμα της έκφρασης των γονιδίων μπορεί να διαφοροποιείται ,από το βήμα της μεταγραφής DNA-RNA στη μετα-μεταγραφική τροποποίηση της πρωτεΐνης. Η σταθερότητα του τελικού γονιδιακού πρϊόντος , είτε είναι πρωτεΐνη έιτε RNA ,συνεισφέρει επίσης στο επίπεδο έκφρασης του γονιδίου. Ένα μη σταθερό γονιδιακό προϊόν έχει ως αποτέλεσμα μια έκφραση χαμηλότερου επιπέδου. Γενικά η γονιδιακή έκφραση ρυθμίζεται μέσω των αλλαγών στον αριθμό και στον τύπο των αλληλεπιδράσεων ,μεταξύ των μορίων τα οποία επηρεάζουν συλλογικά τη μεταγραφή του DNA και την μετάφραση του RNA.

1.5.2 Μεταγραφική ρύθμιση

Η ρύθμιση της μεταγραφής μπορεί να χωριστεί σε τρεις τρόπους που την επηρεάζουν: γενετικό(άμεση ανίδραση ενός ελεγκτικού παράγοντα με το γονίδιο), διαφοροποίηση (αντίδραση ενός παράγοντα ελέγχου με την μηχανή μεταγραφής) και επιγενετικός (μη ακολουθιακές αλλαγές στη δομή οτυ DNA που επηρεάζει τη μεταγραφή).

Η άμεση αλληλεπίδραση με το DNA είναι η πιο απλή και άμεση μέθοδος που μπορεί μια πρωτεΐνη να αλλάξει τα επίπεδα της μεταγραφής και τα γονίδια συχνά έχουν διάφορα σημεία όπου μπορούν να προσδεθούν πρωτεΐνες ,γύρω από την κωδικοποιητική περιοχή ,με τη συγκεκριμένη λειτουργία της ρυθμιστικής μεταγραφής.Υπάρχουν πολλά ρυθμιστικά σημεία πρόσδεση στο DNA γνωστά ως ενισχυτές ,μονωτήρες ,καταστολείς και αποσιωπητές. Οι μηχανισμοί της ρυθμιστικής μεταγραφής διαφέρουν πολύ, από το μπλοκάρισμα των βασικών σημείων πρόσδεσης στο DNA για την RNA πολυμεράση ,μέχρι την δράση ως ενεργοποιητής και προωθητής της μεταγραφής βοηθώντας την RNA πολυμεράση να συνδεθεί.

Η δραστηριότητα των μεταγραφικών παραγόντων ,διαμορφώνεται στη συνέχεια από ενδοκυτταρικά σήματα δημιουργώντας μετα-μεταγραφική τροποποίηση της πρωτεΐνης. Το γεγονός αυτό επηρεάζει την ικανότητα του μεταγραφικού παράγοντα να προσδεθεί ,έμμεσα ή άμεσα ,στον προωθητή DNA ,την πρόσληψη της RNA πολυμεράσης ,ή να ευνοήσει την την επιμήκυνση ενός νέου συντιθεμένου μορίου RNA.

Η μεμβράνη του πυρήνα στα ευκαρυωτικά κύτταρα επιτρέπει περαιτέρω ρύθμιση των μεταγραφικών παραγόντων ,μέσω της διάρκειας της παρουσίας τους στον πυρήνα ,η οποία ρυθμίζεται από αναστρέψιμες αλλαγές στη δομή τους και μέω της σύνδεσης άλλων πρωτεϊνών.Περιβαλλοντικά ερεθίσματα ή ενδοκρινολογικά σήματα μπορεί να προκαλέσουν αλλαγές στις ρυθμιστικές πρωτεΐνες προκαλώντας αλυσιδωτές αντιδράσεις ενδοκυτταρικών σημάτων, που έχουν αποτέλεσμα στην ρύθμιση της γονιδιακής έκφρασης.

Πρόσφατα έχει γίνει φανερό το γεγονός της τεράστιας επιρροής μη DNA ακολουθιών με συγκεκριμένα αποτελέσματα στη μετάφραση. Τα αποτελέσματα αυτά αναφέρονται ως επιγενετικά και συνεπάγονται τη μεγαλύτερη δομή του DNA, τις μη ακολουθιακές ειδικές DNA πρωτεΐνες πρόσδεσης και τη χημική μεταβολή του DNA. Σε γενικές γραμμές οι επιγενετικές επιδράσεις αλλοιώνουν την προσβασιμότητα του DNA στις πρωτεΐνες και έτσι διαμορφώνουν τη μεταγραφή.

Η μεθυλίωση του DNA είναι ένας ευρέως διαδεδομένος μηχανισμός για επιγενετική επίδραση στη γονιδιακή έκφραση και έχει παρατηρηθεί σε βακτήρια και έχει ρόλους στη σίγαση της κληρονομικής μεταγραφής και στη ρύθμιση της μεταγραφής.Στους ευκαρυωτικούς οργανισμούς ,η δομή της χρωματίνης ,ρυθμίζει την πρόσβαση στο DNA με σημαντικές επιπτώσεις στην έκφραση των γονιδίων στις περιχές ευμοχρωματίνης και ετεροχρωματίνης.

1.5.3 Μετα-μεταγραφική ρύθμιση

Στα ευκαρυωτικά κύτταρα, όπου η εξαγωγή RNA είναι απαραίτητη πριν τη μετάφραση και είναι δυνατή, η πυρηνική εξαγωγή θεωρείται ότι παρέχει επιπλέον έλεγχο πάνω στη γονιδιακά έκφραση. Όλη η μεταφορά μέσα και έξω από τον πυρήνα ,γίνεται μέσω του πυρηνικού πόρου και η μεταφορά ελέγχεται μέσω μιας μεγάλης ποκιλίας πρωτεινών importin και exportin.

Η έκφραση ενός γονιδίου που κωδικοποιεί για μια πρωτεΐνη είναι δυνατή μόνο αν το αγγελιοφόρο RNA που μεταφέρει τον κώδικα ,επιβίωσει αρκετά έτσι ώστε να μεταφραστεί. Σε ένα τυπικό κύτταρο ,ένα μόριο RNA είναι σταθερό μόνο αν είναι προσταευμένο από την υποβάθμιση. Η υποβάθμιση του RNA έχει ιδιαίτερη σημασία στη ρύθμιση της έκφρασης στα ευκαρυωτικά κύτταρα, όπου το mRNA έχει να διανύσει σημαντικές αποστασεις πριν μεταφραστεί.Στα ευκαρυωτικά το RNA σταθεροποιείται με συγκεκριμένες μετα-μεταγραφικές μεταβολές.

Η επιτηδευμένη υποβάθμιση των mRNA δεν χρησιμοποιείται μόνο ως μηχανισμός άμυνας των ξένων RNA (συνήθως ιοί RNA), αλλά επίσης ως μια οδός αποσταθεροποίησης του mRNA. Αν ένα mRNA μόριο ,έχει μια συπληρωματική ακολουθία σε ένα μικρό παρεμβατικό RNA, τότε είναι στοχευμένο για καταστροφή μέσω του μονοπατιού της RNA παρεμβολής.

1.5.4 Μεταφραστική ρύθμιση

Η άμεση ρύθμιση της μετάφραση είναι λιγότερο διαδεδομένη από ότι ο έλεγχος της μεταγραφής ή της σταθερότητας του mRNA και χρησιμοποιείται περιστασιακά.Η αναστολή της πρωτεϊνικής μετάφρασης είναι ένας σημαντικός στόχος για τις τοξίνες και τα αντιβοτικά ,με σκοπό να σκοτώσουν ένα κύτταρο επιτάσσοντας τον κανονικό έλεγχο της γονιδιακής έκφρασης. Οι αναστολείς της πρωτεϊνικής σύνθεσης το αντιβιοτικό νεομυκίνη και την τοξίνη ρικίνη.

1.6 Αναπαράσταση DNA ως συμβολοσειρά

Κάθε μόριο DNA να αναπαρασταθεί ως μια ακολουθία συμβόλων (συμβολοσειρά), από ένα το αλφάβητο των τεσσάρων χαρακτήρων {A,T,C,G}.Το A χρησιμοποιείται για την αδενίνη, το T για την θυμίνη, το C για την κυτοσίνη και το G για την γουανίνη.Η υπόθεση αυτή είναι πολύ σημαντική για την υπολογιστική επεξεργασία και αποθήκευση. Ο προσδιορισμός αυτής της συμβολοσειράς για διαφορετικά μόρια ή ο προσδιορισμός της σειράς των συμβόλων βάσεων στα μόρια, είναι ένα κρίσιμο βήμα για την κατανόηση των βιολογικών λειτουργιών των μορίων.Γενικότερα υποθέτουμε πως κάθε βιολογικό μόριο μπορεί να αναπαρασταθεί ως μια ακολουθία συμβόλων από ένα συγκεκριμένο αλφάβητο Σ.

1.7 Μοτίβα και τρόποι εύρεσης τους.

Τα ακολουθιακά μοτίβα είναι μικρά, επαναλαμβανόμενα μοτίβα στο DNA τα οποία εκτιμούνται ότι έχουν βιολογική λειτουργικότητα. Συχνά δείχνουν ειδικές θέσεις πρόσεδης στην ακολουθία για πρωτεΐνες ,όπως οι νουκλεάσες και οι μεταγραφικοί παράγοντες (TF). Άλλα συμμετέχουν σε σημαντικές διεργασίες στο επίπεδο του RNA ,συμπεριλαμβάνοντας τη σύνδεση με το ριβόσωμα ,την επεξεργασία των mRNAs και τον τερματισμό της μεταγραφής.

Όταν ένα μοτίβο ακολουθίας εμφανίζεται στο εξώνιο ενός γονιδίου , μπορεί να κωδικοποιήσει το «δομικό μοτίβο» μιας πρωτεΐνης. Αυτό είναι το στερεοτυπικό στοιχείο ολόκληρης της δομής της πρωτεΐνης. Παρ'όλα αυτά τα μοτίβα δε χρειάζεται να συνδεθούν με μια διακριτική δευτερεύουσα δομή. Οι «μη-κωδικοποιητικές» ακολουθίες δε μεταφράζονται σε πρωτεΐνες και τα νουκλεϊκά οξέα με τέτοια μοτίβα δε χρεάζεται να αποκλίνουν αό την τυπικό σχήμα.

Έξω από τα εξώνια του γονιδίου, υπάρχουν ρυθμιστικά ακολουθιακά μοτίβα και μοτίβα μέσα στην περιοχή "junk" του DNA (κομμάτια του DNA των οποίων η λειτουργία δεν είναι ακόμα γνωστή), όπως το satellite DNA [3]. Μερικά από αυτά θεωρείται ότι επηρεάζουν το σχήμα των νουκλεϊκών οξέων, αλλά είναι κάτι που συμβαίνει απλά μερικές φορές. Για παράδειγμα πολλές πρωτεΐνες που συνδέονται με το DNA που έχουν συγγένεια με συγκεριμένα μοτίβα, συνδέονται με το DNA μόνο όταν είναι στη διπλή ελικοειδή του μορφή. Είναι σε θέση να αναγνωρίζουν μοτίβα με την επαφή τους με μεγάλα ή μικ΄ρα αυλάκια της διπλής έλικας.

Τα μικρά κωδικοποιητικά μοτίβα , τα οποία φαίνεται να έχουν έλλειψη δευτερεύουσας δομής ,περιλαμβάνουν αυτά που ονοματίζουν τις πρωτεΐνες για παράδοσης τους σε ορισμένα μέρη του κυττάρου, ή τα σημειώνουν για φωσφορυλίωση.

Μέσα σε μια ακολουθία ή σε μια βάση δεδομένων με ακολουθίες ,οι ερευνητές ψάχνουν και βρίσκουν μοτίβα χρησιμοποιώντας τεχνικές ανάλυσης των ακολουθιών που βασίζονται στον υπολογιστή ,όπως είναι ο BLAST.

1.7.1 Μοτίβα και ομόφωνες ακολουθίες (consensus sequences)

Στη μοριακή βιολογία και στη βιοπληροφορική μια ομόφωνη ακολουθία είναι ένας τρόπος αναπαράστασης των αποτελεσμάτων μιας πολλαπλής ευθυγράμμισης ακολουθίας , όπου συσχετιζόμενες ακολουθίες συγκρίνονται η μία με την άλλη και βρίσκονται όμοια λειτουργικά ακολουθιακά μοτίβα. Η ακολουθία consensus δείχνει ποια κατάλοιπα είναι περισσότερο άφθονα στην ευθυγράμμιση σε κάθε θέση.

Ο συμβολισμός [XYZ] εννοεί το X, το Y ή το Z , αλλά δεν αναφέρει την πιθανότητα ενός συγκεκριμένου ταιριάσματος. Γι'αυτό το λόγο ,δύο ή περισσότερα μοτίβα είναι συχνά συνδεδεμένα ένα ενίαίο μοτίβο: το καθοριστικό μοτίβο και διάφροα άλλα τυπικά μοτίβα.

Η ανάπτυξη λογισμικού για αναγνώριση μοτίβου αποτελεί μείζον θέμα της γενετικής ,της μοριακής βιολογίας και της βιοπληροφορικής. Συγκεκριμένα ακολουθιακά μοτίβα μπορούν να λειτουργήσουν ως ρυθμιστικές ακολουθίες που ελέγχουν τη βιοσύνθεση ,ή ως ακολουθίες σήματος που κατευθύνουν ένα μόριο σε ένα συγκεκριμένο σημείο στο κύτταρο ή ρυθμίζουν την ωρίμανση του. Εφόσον η ρυθμιστική λειτουργία αυτών τνων ακολουθιών έιναι σημαντική πιστεύεται ότι αυντηρούνται μέσα στο πέρασμα των περιόδων εξέλιξης. Σε μερικές περιπτώσεις ,η εξελικτική συγγένεια μπορεί να εκτιμηθεί από το ποσοστό συντήρησης αυτών των περιοχών.

1.7.2 De novo υπολογιστική ανακάλυψη μοτίβων

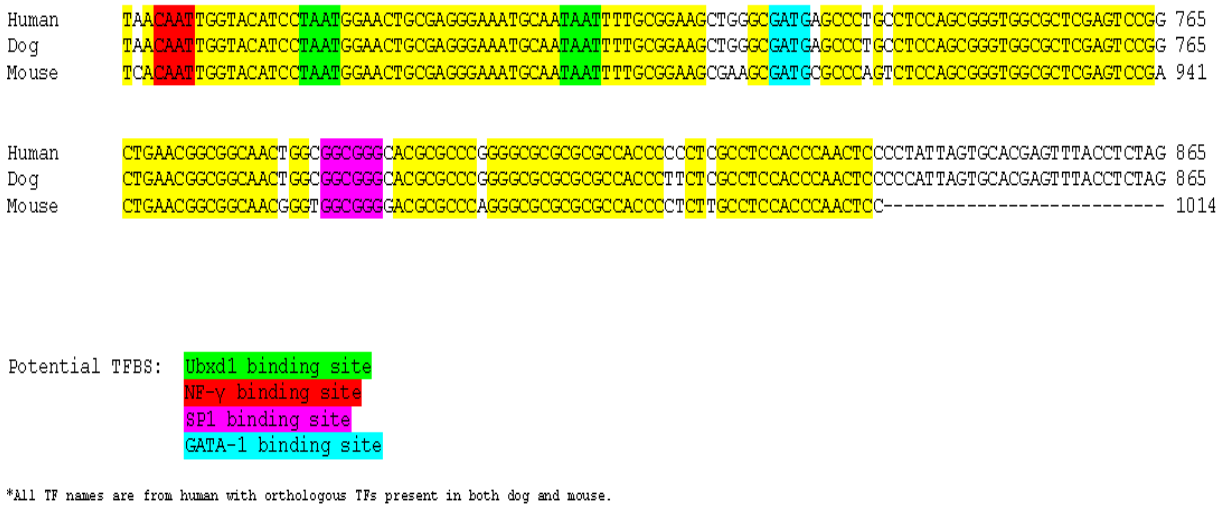
Υπάρχουν προγράμματα λογισμικού τα οποία ,δεδομένων πολλαπλών ακολουθιών ως είσοδο ,προσπαθούν να αναγνωρίσουν ένα ή περισσότερα υποψήφια μοτίβα. Ένα παράδειγμα έιναι το MEME ,το οποίο παράγει στατιστικές πληροφορίες για το κάθε υποψήφιο. Άλλοι αλγόριθμοι περιλαμβάνουν τους CisModule, AlignAce, PhyloGibbs, Weeder, Amadeus και FIRE. Το SCOPE είναι ένας ανιχνευτής συνόλου μοτίβων που χρησιμοποιεί διάφορους αλγόριθμους ταυτόχρονα. Σήμερα υπάρχουν περισσότερες από 100 δημοσιεύσεις με παρόμοιους αλγόριθμους ,χωρίς να υπάρχει μια συνολική συγκριτική αξιολόγηση ,κι έτσι η επιλογή κάποιου δεν είναι τόσο εύκολη υπόθεση.

1.7.3 Ανακάλυψη μέσω εξελικτικής συντήρησης

Τα μοτίβα έχουν ανακαλυφθεί μελετώντας όμοια γονίδια σε διαφορετικά είδη. Για παράδειγμα ,μέσω της στοίχισης ακολουθιών αμινοξέων ορισμένες από το GCM (νευρογλοιακά κύτταρα που λείπουν) γονίδιο ατον άνθρωπο ,στο ποντίκι και στο D.melanogaster,ο Akiyama [4] και άλλοι ανακάλυψαν ένα πρότυπο το οποίο το ονόμασαν το GCM μοτίβο.

Οι συγγραφείς ήταν σε θέση να δείξουν πως το μοτίβο έχει δραστηριότητα DNA σύνδεσης. Ένας αλγόριθμος ανακάλυψης μοτίβου το οποιό λαμβάνει υπόψιν φυλλογενετική συντήρηση, είναι ο PhylloGibs.

Το φυλλογενετικό αποτύπωμα είναι μια τεχνική μέθοδος που χρησιμοποιείται για την αναγνώριση σημείων σύνδεσης μεταγραφικών παραγόντων σε μια κωδικοποιητική περιοχή ενδιαφέροντος του DNA ,συγκρίνοντας την με ορθολογικές ακολουθίες σε διαφορετικά είδη. Αυτό το κάνει αναγνωρίζοντας τα καλύτερα συντηρούμενα μοτίβα σε τέτοιες ορθολογικές περιοχές.



Εικόνα:φυλλογενετικό αποτύπωμα του γονιδίου HOXA5

1.8 Εφαρμογές

Στην ενότητα αυτή θα παρουσιάσουμε ενδεικτικά κάποιες εφαρμογές που έχουν πραγματοποιηθεί για την έυρεση θέσεων πρόσδεσης μεταγραφικών παραγόντων και για την εύρεση θέσεων πρόσδεσης μικρών RNA(MicroRNAs).

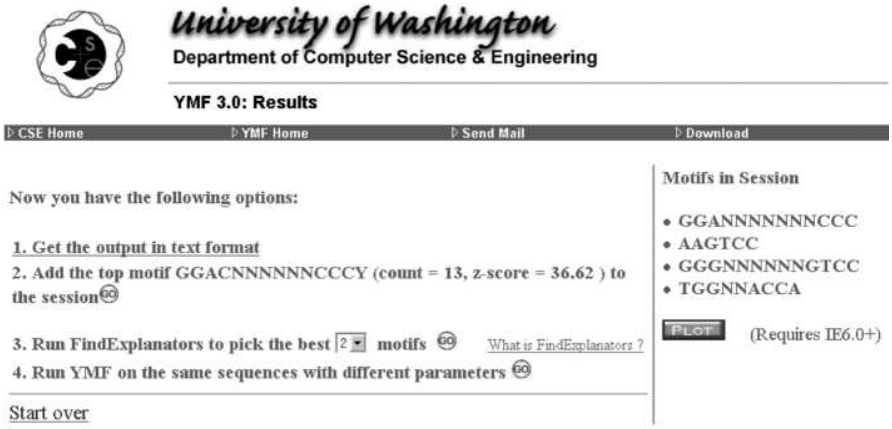
1.8.1 Εύρεση θέσεων πρόσδεσης μεταγραφικών παραγόντων

Μία από τις σημερινές προκλήσεις των βιολόγων είναι η ανακάλυψη νέων περιοχών σύνδεσης στα νουκλεϊκά οξέα για άγνωστους ρυθμιστικούς παράγοντες, δεδομένης μιας συλλογής γονιδίων που πιστεύεται ό τι ρυθμίζονται.

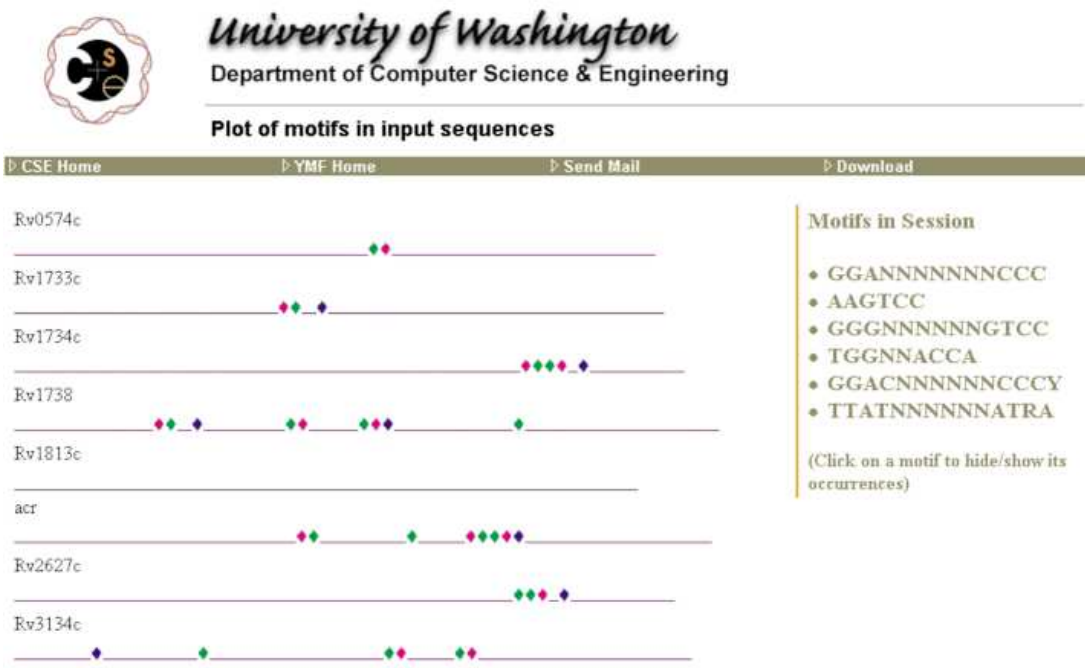
Μια σειρά αλγορίθμων έχουν προταθεί για την εύρεση μοτίβων σε βιολογικές ακολουθίες. Πολλοί από αυτούς στους αλγορίθμους έχουν σχεδιαστεί να βρίσκουν περισσότερο γενικά μοτίβα ,από ότι αυτά που απαιτούνται για την αναγνώριση των περιοχών πρόσεδσης μεταγραφικών παραγόντων.Οι αλγόριθμοι αυτοί βασίζονται σε μεθόδους όπως η expectation maximization [6], Gibs sampling[7][8], και οι άπληστοι αλγόριθμοι[9] ,οποιοσδήποτε από τους οποίους μπορεί να οδηγήσει σε τοπικά βέλτιστη λύση.

YMF

Οι Saurabh Sinha and Martin Tompa [10] ανέπτυξαν ένα πρόγραμμα για την ανακάλυψη νέων σημείων πρόσδεσης για μεταγραφικούς παράγοντες μέσω στατιστικής υπερ-αναπράστασης που ονομάζεται YMF. Ο YMF είναι ένας αλγόριθμος απαρίθμησης ,που δεδομένων των ρυθμιστικών περιοχών διάφορων συσχετιζόμενων γονιδίων ,εγγυάται την παραγωγή μοτίβων με τα μεγαλύτερα z-scores. Το z-score ενός μοτίβου είναι ο αριθμός των τυπικών αποκλίσεων του οποίου ο αριθμός των παρατηρημένων στιγμιοτύπων στις ακριβελις ακολουθίες εισόδου υπερβαίνει τον αναμενόμενο αριθμό στιγμιτοτύπων ,είχε τις ακολουθίες εισόδου αντ'αυτού τυχαίες. Τα μοτίβα μόνα τους είναι μικρές ακολουθίες του IUPAC αλφαβήτου ,με τα N να είναι αναγκαστικά στη μέση της ακολουθίας .



Εικόνα:αρχική σελίδα του εργαλείου YMF



Εικόνα: σχεδίαση του YMF-τα μοτίβα που έχουν σχεδιαστεί είναι GGANNNNNNCCC (πράσινο), AAGTCC (μπλε), and GGACNNNNNNCCCY (ροζ)

Οι Caselle M, Di Cunto F, Provero P [11] ανέπτυξαν μια μέθοδο ,η οποία επίσης βασίζεται στη στατιστική υπερ-αναπαράσταση των ρυθμιστικών μοτίβων. Η μέθοδος αυτή: αρχικά ομαδοποιεί τα γονίδια με βάση τα μοτίβα τα οποία υπεραναπαριστώνται στην upstream περισοχή τους. Στη συνέχεια τα γονίδια που αποκτήθηκαν αναλύονται από την οπτική γωνία της συρρύθμισης των γονιδίων που περιέχουν, μελετώντας τα μοτίβα έκφρασης τους σε πειράματα μικροσυστοιχιών. Τα υπερ-αναπαριστώμενα μοτίβα που έχουν τις ετικέτες των ομάδων που δείχνουν ένδειξη

συρρύθμισης ,θεωρούνται ως υποψήφια σημελια πρόσδεσης για τη ρύθμιση των γονιδίων της ομάδας.

Αργότερα οι Davide Cor`a, Ferdinando Di Cunto, Paolo Provero , Lorenzo Silengo and Michele Caselle[12] δοκίμασαν την ίδια μέθοδο από ένα σημείο οπτικής γωνίας σύμφωνα με την οποία , αν ένα σημαντικό τμήμα της ομάδας των γονιδίων που αντιστοιχεί σε ένα δεδομένο μοτίβο, μοιράζονται τον ίδιο λειτουργικό χαρακτηρισμό, τότε είναι πιθανό το μοτίβο αυτό να εμπλέκεται στην συρρύθμιση τους. Η διαδικασία αυτή για την επικύρωση των συνόλων είναι πιθανό να μπορεί να αναγνωρίζει διαφορετικά μοτίβα συκρινόμενα με την επικύρωση από δεδομένα μικροσυστοιχιών ,που περιορίζεται από δύο τουλάχιστον παράγοντες: πρώτον ,οι συγκεκριμένες βιολογικές διαδικασίες και οι περιβαλλοντικές συνθήκες που επικρατούν σε κάθε πείραμα των μικροσυστοιχιών ,και δεύτερον το γεγονός ότι μόνο αρκετά μεγάλες ομάδες συρρυθμιζόμενων γονιδίων είναι πιθανό να παράξουν σημαντικό στατιστικά σήμα.

REDUCE

Οι Crispin Roven και Harmen J.Bussemaker [13] ανέπτυξαν το εργαλείο REDUCE για να συνάξουν ρυθμιστικά στοιχεία από δεδομένα μικροσυστοιχιών. Το REDUCE είναι μια μέθοδς παλινδρόμησης που βασίζεται σε μοτίβα για την ανάλυση μικροσυστοιχιών. Οι μόνες απαιτούμενες είσοδοι είναι μια ενιαία ομάδα γονιδιωμάτων απόλυτης ή σχετικής πληρότητας mRNAs και η ακολουθία DNA της ρυθμιστικής περιοχής που συνδέεται με κάθε γονίδιο που εξετάζεται-ερωτάται. Προς το παρόν υποστηρίζονται οργανισμοί όπως τα σκουλήκια και οι μύγες και διερωτάται το γεγονός αν μπορεί να υποστηρίξει και οργανισμούς όπως το ποτνίκι και ο άνθρωπος. Το REDUCE χρησιμοποιεί αμερόληπτες στατιστικές για να αναγνωρίσει ολιγονουκλεοτιδικά μοτίβα των οποίων η εμφάνιση τους στην ρυθμιστική περιοχή τνω γονιδίων σχετίζεται με το επίπεδο της έκφρασης του mRNA. Η ανάλυση παλινδρόμησης χρησιμοποιείται για να εξαχθεί η ιδιότητα της μεταγραφικής ενότητας που συνδέεται με κάθε μοτίβο.

Παρακάτω ακολουθεί ένας πίνακας με διάφορες υπολογιστικές μεθόδους και εργαλεία για την ανίχνευση ρυθμιστικών περιοχών πρόσδεσης.[14]

Table 1 Details about the operation principles, basic technical data and URLs of 13 analyzed tools

Program	Operating principle	Technical data	URL	Reference
AlignACE	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are over-represented in the input set	Judges alignments sampled during the course of the algorithm using a maximum *a priori* log likelihood score, which gauges the degree of overrepresentation. Provides an adjunct measure (group specificity score) that takes into account the sequence of the entire genome and highlights those motifs found preferentially in association with the genes under consideration.	http://atlas.med.harvard.edu/	7
ANN-Spec	Models the DNA-binding specificity of a transcription factor using a weight matrix	Objective function based on log likelihood that transcription factor binds at least once in each sequence of the positive training data compared with the number of times it is estimated to bind in the background training data. Parameter fitting is accomplished with a gradient descent method, which includes Gibbs sampling of the positive training examples.	http://www.cbs.dtu.dk/~workman/ann-spec/	8
Consensus	Models motifs using weight matrices, searching for the matrix with maximum information content	Uses a greedy method, first finding the pair of sequences that share the motif with greatest information content, then finding the third sequence that can be added to the motif resulting in greatest information content, and so on.	http://bifrost.wustl.edu/consensus/	9
GLAM	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output	Since the basic algorithm cannot find multiple motif instances per sequence, long sequences were fragmented into shorter ones, and the alignment was transformed into a weight matrix and used to scan the sequences to obtain the final site predictions.	http://zlab.bu.edu/glam/	10
The Improbizer	Uses expectation maximization to determine weight matrices of DNA motifs that occur improbably often in the input sequences	As a background (null) model it uses up to a second-order Markov model of background sequence. Optionally, Improbizer constructs a Gaussian model of motif placement, so that motifs that occur in similar positions in the input sequences are more likely to be found.	http://www.soe.ucsc.edu/~kent/improbizer	11
MEME	Optimizes the E-value of a statistic related to the information content of the motif	Rather than sum of information content of each motif column, statistic used is the product of the *P* values of column information contents. The motif search consists of performing expectation maximization from starting points derived from each subsequence occurring in the input sequences. MEME differs from MEME3 mainly in using a correction factor to improve the accuracy of the objective function.	http://meme.sdsc.edu/	12
MITRA	Uses an efficient data structure to traverse the space of IUPAC patterns.	For each pattern, MITRA computes the hypergeometric score of the occurrences in the target sequences relative to the background sequences and reports the highest scoring patterns.	http://www.calit2.net/compbio/mitra/	13
MotifSampler	Matrix-based, motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model	The probabilistic framework is further exploited to estimate the expected number of motif instances in the sequence.	http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html	14

Table 1 Continued

Program	Operating principle	Technical data	URL	Reference
Oligo/dyad-analysis	Detects overrepresented oligonucleotides with oligo-analysis[15] and spaced motifs with dyad-analysis[16]	These algorithms detect statistically significant motifs by counting the number of occurrences of each word or dyad and comparing these with expectation. Most crucial parameter is choice of appropriate probabilistic model for the estimation of occurrence significance. In this study, a negative binomial distribution on word distributions was obtained from 1,000 random promoter selections of the same size as the test sets	http://rsat.scmbb.ulb.ac.be/rsat/	15,16
QuickScore	Based on an exhaustive searching algorithm that estimates probabilities of rare or frequent words in genomic texts	Incorporates an extended consensus method allowing well-defined mismatches and uses mathematical expressions for efficiently computing z-scores and *P* values, depending on the statistical models used in their range of applicability. Special attention is paid to the drawbacks of numerical instability. The background model is Markovian, with order up to 3.	http://algo.inria.fr/dolley/QuickScore/	17
SeSiMCMC	Modification of Gibbs sampler algorithm that models the motif as a weight matrix, optionally with the symmetry of a palindrome or of a direct repeat, and optionally with spacers	Includes two alternating stages. The first one optimizes the weight matrix for a given motif and spacer length. The algorithm changes the positions of the motif occurrences in the sequences and infers the motif model from the current occurrences. These changes are used to optimize the likelihood of sequences as being segmented into the (Bernoulli) background and the motif occurrences. The optimization is organized via a Gibbs-like Markov chain, which samples positions in sequences one by one, until the Markov chain converges. The second stage looks for best motif and spacer lengths for obtained motif positions. It optimizes the common information content of motif and of distributions of motif occurrence positions.	http://favorov.hole.ru/gibbslfm/	18
Weeder	Consensus-based method that enumerates exhaustively all the oligos up to a maximum length and collects their occurrences (with substitutions) from input sequences	Each motif evaluated according to number of sequences in which it appears and how well conserved it is in each sequence, with respect to expected values derived from the oligo frequency analysis of all the available upstream sequences of the same organism. Different combinations of 'canonical' motif parameters derived from the analysis of known instances of yeast transcription factor binding sites (length ranging from 6 to 12, number of substitutions from 1 to 4) are automatically tried by the algorithm in different runs. It also analyzes and compares the top-scoring motifs of each run with a simple clustering method to detect which ones could be more likely to correspond to transcription factor binding sites. Best instances of each motif are selected from sequences using a weight matrix built with sites found by consensus-based algorithm.	http://159.149.109.16/Tool/ind.php	19
YMF	Uses an exhaustive search algorithm to find motifs with the greatest *z*-scores	A *P* value for the z-score is used to assess significance of motif. Motifs themselves are short sequences over the IUPAC alphabet, with spacers ('N's) constrained to occur in the middle of the sequence.	http://bio.cs.washington.edu/software.html#ymf	20

1.8.2 Εύρεση θέσεων πρόσδεση microRNAS

Αυξανόμενες ενδείξεις δείχνει πως οι 3’ μη μεταφρασμενες περιοχές (3’ UTRs) των mRNAs περιέχουν διαφορετικούς τύποθς μικρών ακολουθιακών στοιχείων που παίζουν ένα σημαντικό ρόλο στο μετα-μεταγραφικό έλεγχο της γονιδιακής έκφρασης ,ρυθμίζοντας την σταθερότητα του mRNA ,την θέση του και την αποτελεσμάτικοτηα της μετάφρασης.

microRNAs

Τα MicroRNAs είναι μια κατηγορία μικρών RNAs που δεν κωδικοποιούν πρωτεϊνες ,τα οποία θεωρούνται ότι είναι σημαντικά σε πολλές βιολογικές διαδικασίες μέσω της ρύθμισης της γονιδιακής έκφρασης.

Από την ανακάλυψη του πρώτου miRNA ,του Lin 4, μια δεκαετία πριν περίπου, πάνω από 3500 μοναδικά miRNAs έχουν ταυτοποιηθεί και έχουν καταχωρηθεί σε ένα μητρώο miRNA (http://microrna.sanger.ac.uk). Μεταξύ αυτών περισσότερα από 300 miRNAs έχουν ανακαλυφθεί στους ανθρώπους μέχρι σήμερα.

Κάθε miRNA ρυθμίζει την έκφραση ενός μεγάλου αριθμού στόχων-γονιδίων ,ανστέλλοντας της μετάφραση της πρωτεϊνης –στόχου,ενώ πειραματικά στοιχεία δείχνουν ότι από ένα μόνο miRNA , μπορούν να ρυθμιστούν έως 200 γονίδια.Συνεπώς τα μόρια αυτά φαίνεται να ρυθμίζουν το ένα τρίτο όλων των ανθρώπινων γονιδίων κατά το μετα-μεταγραφικό στάδιο καταστέλλοντας το αγγελιαφόρο RNA.Παρόλο που το ανθρώπινο γονιδίωμα περιέχει εκατοντάδες miRNAs (πολλά από τα οποία ίσως δεν είναι μοναδικά στους ανθρώπους) ,ο ρόλος τους στις φυσιολογικές και παθολογικές διαδικασίες μόλις έχει αρχίσει να φαίνεται.

Το κυτταρικό περιβάλλον του miRNA φαίνεται να είναι μοναδικό σε κάθε τύπο κυττάρου και να συνδέεται με διακριτές και πολύ συγκεκριμένες διαδικασίες.Για παράδειγμα το miRNA μπορεί να ρυθμίζει την διαφοροποίηση και τη διατήρηση της ταυτότητας των κυττάρων στο αιμοποιητικό σύστημα,να συμβάλλει στη δημιουργία μυικών φαινοτύπων,να ελέγχει την μορφογένεση των επιθηλιακών ιστών και να ρυθμίζει πτυχές της οργανογένεσης και των μεταβολικών διαδιακασίων.Επίσης οι δομές αυτές ίσως να είναι σημαντικές για τις έμφυτες ανοσολογικές αντιδράσεις του οργανισμού , οι οποίες ελέγχουν ιογενείς λοιμώξεις αναστέλλοντας τη σύνθεης των ιογενών πρωτεϊνών.Τα ιογενή γονιδιώματα επίσης κωδικοποιούν miRNAs ,τα οποία φαίνονται να εξελίσσονται ταχύτατα και να ρυθμίζουν τόσο τον ιογενή κύκλο ζωής όσο και την αλληλεπίδραης μεταξύ των ιών και των φορέων τους.

Για αυτό τον λόγο φαίνεται πως τα μικρά RNAs εμπλέκονται σημαντικά στο φάσμα των βιολογικών μονοπατιών.Μη ομαλές υπογραφές –ίχνη miRNA ίσως υπάρχουν σε καταστάσεις αθένειας και έιναι πολύτιμοι δείκτες διάγνωσης ή/και πρόγνωσης.Θα μπορούσαν επίσης να χρησιμοποιηθούν για τον εντοπισμό ατόμων που βρίσκονται σε κίνδυνοκαι να ελιναι ενδιεκτικά για αλλοιωμένα γενετικά προγράμματα που οδηγούν στην ευαισθησία και την παρουσίαση της ασθένειας.Επιπλέον η διαφοροποίηση των δραστηριοτήτων τους θα μπορούσε να έχει θεραπευτικά οφέλη.Η ταυτοποίηση ορισμένων miRNA υπογραφών – ιχνών έχει επιτευχθεί σε μερικές μορφές καρκίνου και αναμένουμε και για ασθένεις του αναπνευστικού.

Βιογένεση των MicroRNAs

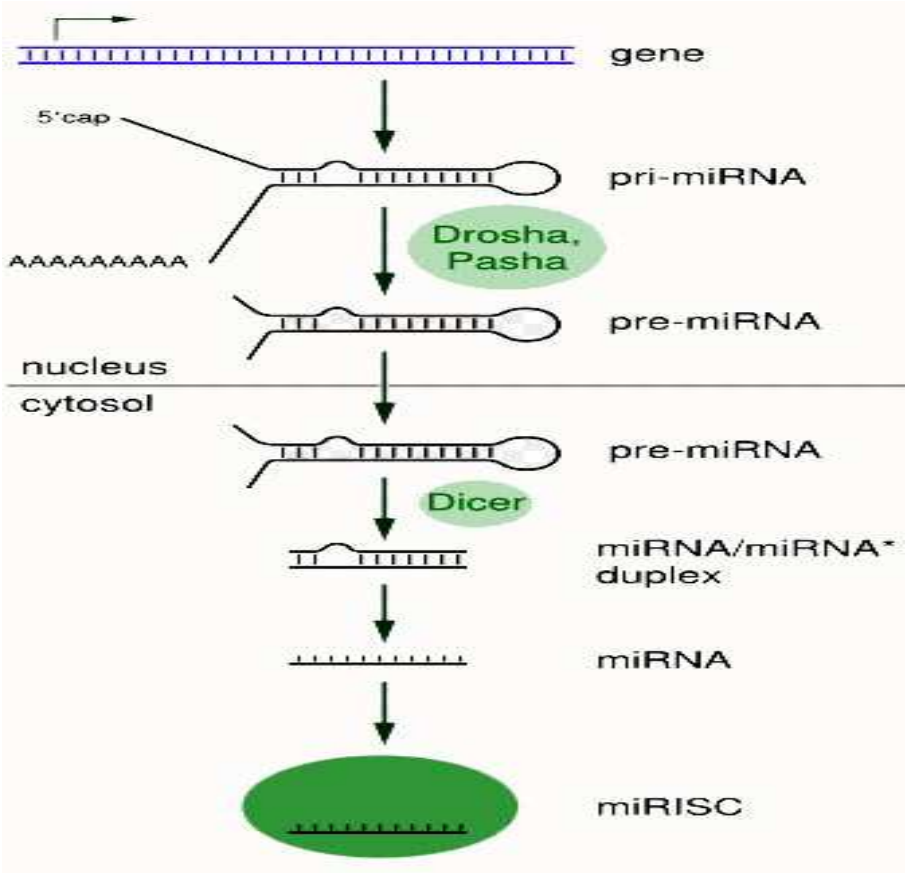
Για την βιογένεση των miRNAs [15] και για τη μεσολάβηση στην γονιδιακή έκφραση απαιτείται ένα πολύπλοκο σύνολο πρωτεϊνών. Τα ολοκληρωμένα μόρια miRNA δημιουργούνται από μεγάλα πρωτογενή μόρια miRNA ,τα pri-miRNA,τα οποία συνήθως περιέχουν εκατοντάδες ζεύγη βάσεων.Τα pri-miRNA επεξεργάζονται στον πυρήνα των "stem-loop precursors" (pre-miRNA) των 70 περίπου νουκλεοτιδίων τις RNase II ενδονουκλεάσες, Drosha και Pasha (20-22).Τα pre-miRNAs μεταφέρονται άμεσα στο κυτταρόπλασμα από πρωτεϊνες ,οι οποίες συμμετέχουν άμεσα στην μεταφορά υλικού του πυρήνα όπως είναι η exportin 5 και Ran-GTP.Στο κυτταρόπλασμα αργότερα μπορούν να μεταβληθούν σε διπλά RNA των 22 νουκλεοτιδίων από την Dicer RNase III και από την διπλού έλικα πρωτεϊνη για τη σύνδεση με το RNA.Η λειτουργική έλικας του miRNA αποκόπτεται από την συμπληρωματική της μη λειτουργική έλικα και εγκαθίσταται στο RISC(RNA-inducing-silencing complex),το οποίο αποτελείται από το Dicer,TRBP και την Argonaute 2 πρωτεϊνη.

Το παραπάνω σύμπλεγμα (complex) ,είναι αυτό που παίζει βασικό ρόλο στη σύνδεση του αγγελιαφόρου RNA(mRNA) μέσω του (pre)-miRNA.Το συγκεκριμένο miRNA-RISC σύνολο στη συγγενές του αγγελιαφόρο RNA-στόχο μέσω μια μικρής περιοχής στο τέλος της ακολουθίας.(5').Η σύνδεση είναι τις περισσότερες φορές ατελής και μερικές βάσεις του αγγελιαφόρου RNA παραμένουν ελεύθερες και δεν συνδέονται στο σύνολο RISC-miRNA.Στη συνέχεια λοιπόν μπορεί η περιοχή 3' να «κουμπώσει» με τις ελεύθερες περιοχές 5' της των αγγελιαφόρων RNA-στόχων.

Σε γενικές γραμμές τα μικρά RNAs μπορούν να λειτουργήσουν με δύο μηχανισμούς:μέσω της διάσπασης του αγγελιαφόρου RNA ή μέσω της εμπόδισης της διαδικασίας της μετάφρασης της κωδικοποιημένης πρωτεϊνης.Η ολοκληρωμένη σύνδεση του μικρού RNA οδηγεί σε διάσπαση ενός ενιαίου φωσφοδιεστερικού bond στο αγγελιαφόρο RNA ακριβώς απέναντι αποτα νουκλεοτίδια 11 και 12.Ακόμα και αν η σύνδεση είναι εντελώς συμπηρωματική , η διάσπαση συμβαίνει όταν το τέλος της περιοχής 5' του μικρού RNA συνδέεται στην Argonaute πρωτεϊνη που φέρει μια ενδονουκλεάση στον χώρο Piwi.Αυτός ο τρόπος είναι γνωστός ως παρέμβαση RNA και είναι ο κύριος μηχανισμός με τον οποίο τα miRNAs των φυτών ρυθμίζουν τη γονιδιακή έκφραση.Στα ζώα ,ωστόσο,το ώριμο miRNA είναι μερικώς συμπηρωματικό στην ακολουθία των αντίστοιχων τους αγγελιαφόρων RNAs,επιτρέποντας τους να συνδέονται με άλλα αγγελιαφόρα RNAs.Αντί να προκαλούν διάσπαση των αγγελιαφόρων RNAs ,τα ζωϊκά miRNAs εμποδίζουν την μετάφραση της κωδικοποιημένης πρωτεϊνης, υπογραμμίζοντας την δυνητική επιρροή τους σχεδόν σε κάθε γενετικό μηχανισμό και την πιθανή τους συμβολή σε ασθένειες.Ωστόσο η σύνδεση ενός miRNA μπορεί να μην είναι επαρκής να μπλοκάρει τη μετάφραση και ίσως να απαιτούνται διάφορες άλλες ομάδες miRNAs να συνδεθούν με το αγγελιαφόρο RNA για τον συνδυαστικό έλεγχο της γονιδιακής έκφρασης.

Αρχικά θεωρούνταν πως η σύνδεση του miRNA στο RISC ανέστελε την σύνθεση της πρωτεϊνης υποβαθμίζοντας την νέα σύνθεση της όπως προέκυπτε από το ριβόσωμα,ή παγώνοντας τα ριβοσώματα σε μέρος του αγγελιαφόρου RNA και έτσι μπλόκαρε την επιμήκυνση της πρωτεϊνης.Νέα στοιχεία ωστόσο δείχνουν πως το miRNA μπορεί να προκαλέσει αστάθεια στην στον στόχο.Το miRNA-RISC σύνολο όταν είναι συνδεδεμένο με τον στόχο του εμποδίζει την

έναρξη της μετάφρασης προωθώντας την κίνηση του αγγελιαφόρου RNA από το κυτταρόπλασμα σε μέρη του RNA,που ονομάζονται P-bodies.Γι'αυτό τον λόγο τα miRNAs αναστέλλουν δυνητικά την έναρξη της μετάφρασης των mRNA στις πρωτεϊνες ,προωθώντας την εγκατάσταση των mRNAs σε χώρους καταστροφής του RNA.



Αρκετές υπολογιστικές προσεγγίσεις έχουν αναπτυχθεί τα τελευταία χρόνια για την διερεύνηση του ρυθμιστισκού μηχανισμού των miRNAs. Συγκεκριμένα προτάθηκαν προσεγγίσεις για τα ακόλουθα προβλήματα:

- Αναγνώριση των miRNA γονιδίων
- Αναγνώριση των γονιδίων που ρυθμίζονται από τα miRNAs
- Περιγραφή του ρυθμιστικού δικτύου που καθιερώνεται από αυτή την τάξη μορίων

Οι περισσότερες υπολογιστικές μέθοδοι που προτάθηκαν για την αναγνώριση των στόχων των miRNAs βασίζονται σε μερικά από τα ακόλουθα στοιχεία:

- Εξελικτική συντήρηση των miRNAs και των σημείων πρόσδεσης τους ανάμεσα στα είδη.
- Χρήση του Watson-Crick τέλειων ή μη ζευγαριών ανάμεσα στα 3' UTRs και στα miRNAs.

- Εμπλουτισμό των σημείων πρόσδεσης των miRNAs στις 3’UTRs
- Χρήση της δευτερεύουσας δομής RNA

Οι Davide Cor`a , Ferdinando Di Cunto , Michele Caselle and Paolo Provero[16] ανέπτυξαν μια εφαρμογή για τον εντοπισμό υποψήφιων ρυθμιστικών ακολουθιών στις 3’UTRs των θηλαστικών με τη στατιστική ανάλυση των ολιγονουκλεοτιδικών κατανομών.

Συγκεκριμένα παρουσίασαν δύο νέες μεθόδους για τον εντοπισμό περιοχών σύνδεης των miRNAs ή γενικότερα ρυθμιστικών ακολουθιών στις 3’UTRs που βρίσκοντια στο mRNA.Οι μέθοδοι βασίζονται στη συχνότητα κατανομής των ολιγονουκλεοτιδίων στις 3’UTRs και στην εξελικτική συντήρηση και συγκεκριμένα στις παρακάτω δύο υποθέσεις:

1. *Συντηρούμενη υπερεκπροσώπηση*: τα σημεία πρόσδεσης συμβαίνουν στα ρυθμισμένα γονίδια πιο συχνά από ότι θα περιμέναμε να συμβεί τυχαία,και αυτή η υπεραναπαράσταση μπορέι να διατηρηθεί σε ορθόλογα γονίδια συγγενικών ειδών.
2. *Ασσυμετρία σύνδεσης*: τα σημεία σύνδεσης εμφανίζονται στις 3’ UTRs πιο συχνά από ότι στο αντίστροφο συμπλήρωματικό τους. Το σκεπτικό είναι ότι αν πολλά γονίδια είναι υπό θετική επιλεκτική πίεση να διατηρήσουν τις περιοχές σύνδεσης στις 3’UTRs τους, αυτό θα έπρεπε να προκλέσει μια καθολική υπεραναπαράστση των περιοχ΄ν πρόσδεσης συγκριτικά με τα αντίστροφα συμπληρωματικά τους που δεν υπόκεινται σε τέτοια θετική πίεση.

Η κύρια καινοτομία της μεθόδου διατηρητέας υπεραναπαράστασης είναι ότι και το μήκος των 3’UTRs και της καθολικής νουκλεοτιδικής σύνθεσης τους ,λαμβάνεται υπόψιν όταν προσδιορίζουμε αν ένα συγκεκριμένο ολιγονουκλεοτίδιο υπεραναπαριστάται σε μια δεδομένη 3’UTR.

ΚΕΦΑΛΑΙΟ 2

Μέθοδοι

2.1 Περιγραφή της διαδικτυακής εφαρμογής

Στην ενότητα αυτή θα περιγράψουμε τη διαδικτυακή εφαρμογή και πώς αυτή λειτουργεί.

Στόχος της συγκεκριμένης εφαρμογής είναι η σχεδίαση και η ανάπτυξη ενός ολοκληρωμένου διαδικτυακού περιβάλλοντος , στο οποίο ο κάθε χρήστης θα μπορεί να κάνει upload αρχεία που θα περιέχουν ακολουθίες DNA ,που ανήκουν σε δύο διαφορετικές ομάδες δεδομένων για επεξεργασία.

Αρχικά λοιπόν ο χρήστης θα πρέπει να κάνει εγγραφή στο σύστημα για να μπορέσει να κάνει upload τα αρχεία του. Δίνοντας λοιπόν το email του ως όνομα χρήστη ,αυτόματα λαμβάνει ένα email από το σύστημα όπου τον καλωσορίζει στην εφαρμογή και του δίνει το password για να κάνει login. Το password παράγεται από μια τυχαία γεννήτρια αλφαριθμητικών μήκους 8 χαρακτήρων. Ταυτόχρονα γίνεται επικοινωνία με τη βάση δεδομένων για εισαγωγή των στοιχείων του χρήστη.

Σε δεύτερη φάση ο χρήστης πραγματοποιεί την είσοδο στο σύστημα για να μπορέσει να κάνει upload αρχεία του.Γίνεται επικοινωνία με τη βάση δεδομένων για την επαλήθευση των στοιχείων του χρήστη ,ώστε να μπορέσει να εισέλθει στο σύστημα.Σε περίπτωση λάναθασμένων στοιχείων ,ο χρήστης μεταφέρεται αυτόματα σε μια σελίδα όπου τον ενημερώνει να ελέγξει ξανά τα στοιχεία εισαγωγής και να προσπαθήσει να κάνει ξανά login.

Μετά την επιστυχημένη είσοδο του χρήστη στο σύστημα, ο χρήστης μπορεί πλέον να δώσει τα αρχεία του για επεξεργασία. Αρχικά κάνει upload το αρχείο που περιέχει τις ακολουθίες UP. Αφού το ανεβάσει στον server μεταφέρεται στη σελίδα όπου του ζητάει να κάνει upload και το δεύτερο αρχείο που περιέχει τις ακολουθίες OTHER. Αυτό γίνεται για να είναι ξεκάθαρο στον χρήστη η σειρά με την οποία πρέπει να ανεβάσει τα αρχεία ,ώστε να γίνει σωστή επεξεργασία των δεδομένων και να λάβει τα ανάλογα σωστά αποτελέσματα.

Τα αρχεία UP και OTHER είναι txt αρχεία και τα δεδομένα τους πρέπει να είναι σε FASTA μορφή. Μια ακολουθία σε FASTA μορφή αναπαριστάται ως μια σειρά γραμμών ,οι οποίες δεν πρέπει να ξεπερνάνε τους 120 χαρακτήρες και να μην ξεπερνάνε στη διαδοχή τους 80 χαρακτήρες .Αυτό πιθανόν συμβαίνει για να δεσμεύσουν από πριν τα καθορισμένα μεγάθη των γραμμών στο λογισμικό.

Η πρώτη γραμμή σε ένα αρχείο FASTA μορφής ξεκινάει ή με «>»,ή με ερωτηματικό «;» και θεωρήθηκε σαν ερωτηματικό.Οι υποακόλουθες γραμμές που ξεκινάνε με ερωτηματικό θα πρέπει να αγννοούνται από το λογισμικό.

Η ακολουθία σε FASTA format αρχίζει με μια περιγραφή μιας γραμμής ,που ακολουθείται από τις γραμμές των δεδομένων ακολουθίας .Η γραμμή περιγραφής διακρίνεται από τα δεδομένα της κολουθίας όπως είπαμε και παραπάνω με τη χρήση του συμβόλου (>) στην πρώτη γραμμή.Η λέξη που ακολουθεί μετα το «>» είναι το αναγνωριστικό όνομα της ακολουθίας και το υπόλοιπο της γραμμής είναι η περιγραφή.(και τα δύο είναι προαιρετικά). Δε θα πρέπει να υπάρχουν κενά μεταξύ του «>» και του πρώτου γράμματος του αναγνωριστικού.Συνίσταται όλες οι γραμμές του κειμένου να είμαι μικρότερες από τους 80 χαρακτήρες. Η ακολουθία τελειώνει αν κάποια άλλη γραμμή που ξεκινάει με «>» εμφανιστεί. Αυτό υποδηλώνει την αρχή επόμενης ακολουθίας. Παρακάτω φαίνεται ένα απλό παράδειγμα ακολουθίας σε FASTA format:

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYI
GTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPF
HPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNK
LGGVLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQ
MASILYFSIILAFLPIAGXIENY

Στην εφαρμογή μας σε περίπτωση που στην ακολουθία υπάρχουν σύμβολα όπως «!» ή κενά τα οποία δεν πειτρέπονται τα αφαιρούμε και στη συνέχεια ενημερώνουμε τον χρήστη με ένα log αρχείο ,το οποίο δείχνει σε ποια θέση βρέθηκε ο ανεπιθύμητος χαρακτήρας και πώς αντιμετωπίστηκε.

Στη συνέχεια ένα πρόγραμμα θα κάνει επεξεργασία των ακολουθιών αυτών ως εξής: θα βρίσκει όλα τα non-overlapping εξαμερή στις παραπάνω DNA ακολουθίες.Θα συγκρίνει την συχνότητα εμφάνισης τους , χρησιμοποιώντας ένα έτοιμο στατιστικό τεστ το οποίο αναλύουμε παρακάτω και σύμφωνα με τα αποτελέσματα θα κάνει αναζήτηση στις miRNA ακολουθίες ,ώστε να δει αν κάποιες από αυτές σχετίζονται με τα top scoring εξαμερή.Στο τέλος θα επιστρέφει τα αποτελέσματα με τη μορφή αρχείων στο χρήστη.

2.1.1 Αλγόριθμος για την αναζήτηση των nmers στις ακολουθίες UP και OTHER

Ο αλγόριθμος που χρησιμοποιήθηκε γα την αναζήτηση των nmers(στην περίπτωση μας εξαμερών) στις ακολουθίες UP και OTHER είναι πολύ απλός και λειτουργεί ως εξής:

Παίρνει ως ορίσματα το μήκος των nmers της ακολουθίας προς σύγκριση.στη συνέχεια ψάχνει για την εμφάνιση του nmer στην ακολουθία.Όταν βρει μια εμφάνιση ,συνεχίζει την αναζήτηση από την προηγούμενη θέση αυξημένη κατά n δηλαδή το μήκος του nmer. Διαφορετικά ξεκινάει κάθε φορά από την επόμενη θέση στην ακολουθία.Στο τέλος επιστρέφει για κάθε nmer το normalized_counts το οποίο είναι ίσο με τον αριθμό εμφάνισης του nmer στην ακολουθία προς το μήκος της ακολουθίας UP ή OTHER .

Παράδειγμα

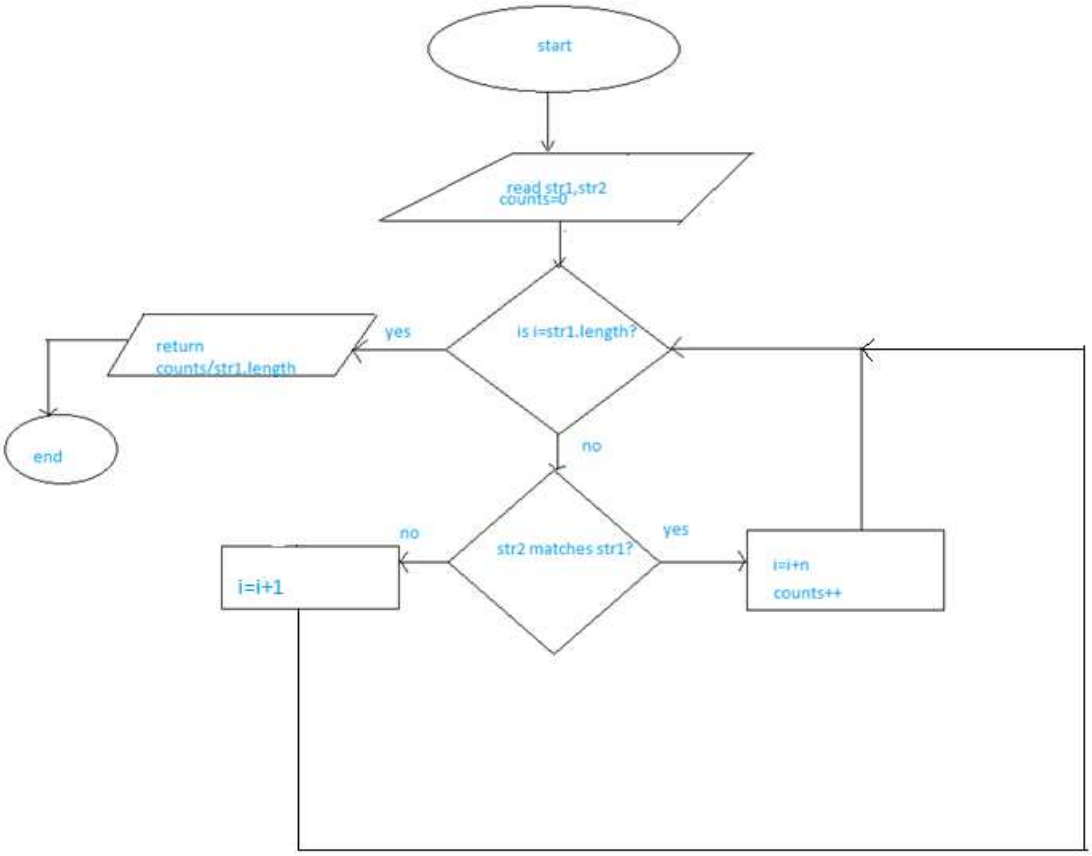
Έστω ότι ψάχνουμε το τριμερές ATA στην ακολουθία CATAGCTATATATTCGATAATAGCT. Το n=3.Σύμφωνα λοιπόν με τον αλγόριθμό μας θα ελέγξουμε αρχικά τις θέσεις 1-3, 2-4 ,5-7 κ.ο.κ

Επομένως θα βρούμε ως στιγιότυπα του ATA στην ακολουθία τα χρωματισμένα με κίτρινο και πράσινο και όχι αυτά με κόκκινο.

CATAGCTATATATTCGATAATAGCT

CATAGCTATATATTCGATAATAGCT

Παρακάτω φαίνεται και μια εικόνα με το διάγραμμα ροής του αλγορίθμου.



Μετά την σύγκριση των ακολουθιών με τα nmers εκτελείται το στατιστικό πακέτο το οποίο υπολογίζει τις τιμές pvalues. Το αρχείο για τη στατιστική επξεργασία παίρνει ως είσοδο δύο αρχεία.Το πρώτο αρχείο περιέχει τα ονόματα των αρχείων που το καθένα από αυτά περιέχει τα αποτελέσματα της σύγκρισης των nmers με τις ακολουθίες των δύο ομάδων ακολουθιών. Σαν έξοδο στο χρήστη θα επιστρέψει ένα αρχείο όπου θα καταγράφονται τα pvalues για κάθε nmer και οι αντίστοιχες –log(pvalues) θα είναι ταξινομημένες από τη μικρότερη προς τη μεγαλύτερη.

Στη συνέχεια επιλέγουμε τα πρώτα top scoring nmers και ελέγχουμε αν ταιριάζουν στις πρώτες θέσεις των MicroRNAs.Στη θέση του U θα πρέοπει να βάλουμε το T.Επειδή όμως πρέπει να τα διάβασουμε από την περιοχή 3’->5’ βρίσκουμε τη συμπληρωματική τους ακολουθία.Έτσι θα μπορέσουμε να δούμε αν υπάρχουν κοινά μοτίβα στα mircRNAs και στα UTRs .Σαν έξοδο ο χρήστης παίρνει ένα αρχείο με τα top scoring nmers και την θέση εμφάνισης του στο κάθε microRNA, το όνομα των MicroRNAs,την αντίστοιχη ln(pvalue) τιμή και μια εικόνα με το ιστόγραμμα των –log(pvalues).

Γενικότερα η διεπαφή της εφαρμογής είναι αρκετά απλή όσον αφορά τα εξωτερικά χαρακτηριστικά, κι αυτό γιατί πρωτεύων στόχος της εφαρμογής είναι η λειτουργικότητα ,κι όχι η εμφανιση της διεπαφής.Η εφαρμογή είναι αρκετά φιλική προς τον χρήστη και αναφέρει σε κάθε σελίδα τα απλά βήματα που πρέπει να κάνει.

2.2 Επιμέρους στοιχεία υλοποίησης

2.2.1 Η Βιοπληροφορική και η γλώσσα προγραμματισμού Java

Η βιοπληροφορική αποτελεί μια πρόκληση για τους προγραμματιστές να δημιουργήσουν ευρέως κατενεμητικά εργαλεία ,τα οποία αποσαφηνίζουν βιολογικές σχέσεις. Η βασική αξία των εργαλείων αυτών , μετράται στην συνεισφορά που παρέχουν στους βιολόγους ερευνητές. Επιπλέον όσο αυξάνεται ο αριθμός και η φύση των διαθέσιμων δεδομένων ,οι επιστήμονες της βιοπληροφορικής κινητοποιούνται ισχυρά να παρέχουν ραγδαία εξελισσόμενα εργαλεία, στις κοινότητες των χρηστών τους.Η Java επέτρεψε στους επιστήμονες της βιοπληροφρικής να αναπτύξουν γρήγορα φιλικές προς το χρήστη εφαρμογές , οι οποίες είναι προσβάσιμες στους χρήστες όλων των επιπέδων υπολογιστικής ικανότητας.

Η ανάπτυξη της βιοπληροφορικής βασισμένη σε Java είναι ένας κλάδος, ο οποίος αναπτύσσεται ταχύτατα. Η Java διευκολύνει την μεταφορά της ακαδημαϊκής κοινότητας, από την ανάπτυξη εφαρμογών που βασίζονται σε απλές script γλώσσες , σε γλώσσες που χειρίζονται πιο δύσκολες λειτουργίες.Οι εφαρμογές αυτές ενσωματώνουν μεγάλο όγκο δεδομένων μια ποικιλία αλγορίθμων που απευθύνονται στην ανάλυση εξειδικευμένων τμημάτων της γονιδιωματικής έρευνας. Η γλώσσα Java αλλάζει επίσης και τον τρόπο με τον οποίο εργάζονται οι επιστήμονες της βιοπληροφορικής και οι βιολόγοι. Με εξειδικευμένα APIs , όπως είναι το JavaHelp , Java 3D, και το Web Services, οι προγραμματιστές μπορούν πολύ γρήγορα να ενσωματώσουν οδηγούς εκμάθησης , να παραδώσουν περισσότερους τύπους δεδομένων και να παρέχουν νέες απόψεις. Με οδηγούμενα APIs ,όπως η BioJava και η EnsEMBL-Java ,οι προγραμματιστές της Java μπορούν πολύ γρήγορα να προσπελάσουν πολύπλοκα βιολογικά αντικείμενα και να τα ενσωματώσουν στις εφαρμογές τους.

2.2.2 Η γλώσσα προγραμματισμού Java

Το περιβάλλον της εφαρμογής υλοποιήθηκε κατά το κύριο μέρος του με τη γλώσσα προγραμματισμού Java.Μια σύντομη περιγραφή των χαρακτηριστικών της συγκεκριμένης γλώσσας ,θα κάνει περισσότερο κατανοητή την επιλογή της.

Η **Java** είναι μία αντικειμενοστρεφής γλώσσα προγραμματισμού που σχεδιάστηκε από την εταιρεία πληροφορικής *Sun Microsystems*.

Μερικά από τα βασικότερα χαρακτηριστικά της Java είναι τα εξής:

- Ένα από τα βασικά πλεονεκτήματα της Java έναντι των περισσότερων άλλων γλωσσών είναι η ανεξαρτησία του λειτουργικού συστήματος και πλατφόρμας. Τα προγράμματα

που είναι γραμμένα σε *Java* τρέχουνε ακριβώς το ίδιο σε Windows, Linux, Unix και Macintosh (σύντομα θα τρέχουν και σε Playstation καθώς και σε άλλες κονσόλες παιχνιδιών) χωρίς να χρειαστεί να ξαναγίνει μεταγλώττιση (compiling) ή να αλλάξει ο πηγαίος κώδικας για κάθε διαφορετικό λειτουργικό σύστημα.

- Η Java χρησιμοποιεί αρχιτεκτονική εικονικής μηχανής (virtual machine), ή (όπως αλλιώς είναι γνωστή) *διερμηνέα* Java (Java Interpreter ή Java runtime). Ο,τιδήποτε θέλει να κάνει ο προγραμματιστής (ή ο χρήστης) γίνεται μέσω της εικονικής μηχανής. Αυτό βοηθάει στο να υπάρχει μεγαλύτερη ασφάλεια στο σύστημα γιατί η εικονική μηχανή είναι υπεύθυνη για την επικοινωνία χρήστη - υπολογιστή. Ο προγραμματιστής δεν μπορεί να γράψει κώδικα ο οποίος θα έχει καταστροφικά αποτελέσματα για τον υπολογιστή γιατί η εικονική μηχανή θα τον ανιχνεύσει και δε θα επιτρέψει να εκτελεστεί. Από την άλλη μεριά ούτε ο χρήστης μπορεί να κατεβάσει «κακό» κώδικα από το δίκτυο και να τον εκτελέσει. Αυτό είναι ιδιαίτερα χρήσιμο για μεγάλα κατανεμημένα συστήματα όπου πολλοί χρήστες χρησιμοποιούν το ίδιο πρόγραμμα συγχρόνως.
- Ακόμα μία ιδέα που βρίσκεται πίσω από τη *Java* είναι η ύπαρξη του **συλλέκτη απορριμμάτων** (*Garbage Collector*). Συλλογή απορριμμάτων είναι μία κοινή ονομασία που χρησιμοποιείται στον τομέα της πληροφορικής για να δηλώσει την ελευθέρωση τμημάτων μνήμης από δεδομένα που δε χρειάζονται και δε χρησιμοποιούνται άλλο. Αυτή η απελευθέρωση μνήμης στη Java είναι αυτόματη και γίνεται μέσω του συλλέκτη απορριμμάτων. Υπεύθυνη για αυτό είναι και πάλι η εικονική μηχανή η οποία μόλις «καταλάβει» ότι ο σωρός (heap) της μνήμης (στη Java η συντριπτική πλειοψηφία των αντικειμένων αποθηκεύονται στο σωρό σε αντίθεση με τη C++ όπου αποθηκεύονται κυρίως στη στοίβα - stack) κοντεύει να γεμίσει ενεργοποιεί το συλλέκτη απορριμμάτων. Έτσι ο προγραμματιστής δε χρειάζεται να ανησυχεί για το πότε και αν θα ελευθερώσει ένα συγκεκριμένο τμήμα της μνήμης, ούτε και για δείκτες (pointers) που αναφέρονται σε άδειο χώρο μνήμης. Αυτό είναι ιδιαίτερα σημαντικό αν σκεφτούμε ότι ένα μεγάλο ποσοστό κατάρρευσης των προγραμμάτων οφείλονται σε λανθασμένο χειρισμό της μνήμης.
- Πολύ καλό documentation.
- Δοκιμασμένη γλώσσα προγραμματισμού για web εφαρμογές.
- Υπάρχει μεγάλο ποσοστό εφαρμογών βιοπληροφορικής που έχουν γραφτεί σε java ή σε scripts, που μπορούν να ενσωματωθούν ή να κληθούν εύκολα από JAVA προγράμματα.
- Οι σημαντικότεροι οργανισμοί παροχής γνώσεως και πληροφορίων στον τομλεα της βιοπληροφορικής, καθώς και οι μεγαλύτερες βάσεις βιολογικών δεδομένων υποστηρίζουν web services με Java clients.
- Η Java είναι πολυνηματική, δηλαδή ένα απλό πρόγραμμα σε Java μπορεί να κάνει πολλά, διαφορετικά προγράμματα ανεξάρτητα και αλληλεπιδρώντα.
- Όλα τα εργαλεία που χρειάζεται κάποιος για να γράψει Java προγράμματα έρχονται δωρεάν, από το περιβάλλον ανάπτυξης μέχρι εργαλεία *build* όπως το Apache Ant και βιβλιοθήκες, ενώ υπάρχουν πολλές διαφορετικές υλοποιήσεις της *Εικονικής Μηχανής* και του *μεταγλωττιστή* (πχ the GNU Compiler for Java) της *Java*. Είναι στο χέρι του καθενός να επιλέξει το κατάλληλο περιβάλλον.
- Για να να γράψει κάποιος κώδικα Java δε χρειάζεται τίποτα άλλο παρά έναν επεξεργαστή κειμένου, όπως το Σημειωματάριο (Notepad) των Windows ή ο vi (γνωστός

στο χώρο του Unix). Παρ'όλ'αυτά, ένα ολοκληρωμένο περιβάλλον ανάπτυξης (*IDE*) βοηθάει πολύ, ιδιαίτερα στον εντοπισμό σφαλμάτων (debugging). Υπάρχουν αρκετά διαθέσιμα, ενώ πολλά από αυτά έρχονται δωρεάν.

2.2.3 Χρήση της JSP

Για την υλοποίηση των δυναμικών σελίδων χρησιμοποιήσαμε πέρα από java και jsp.

Η τεχνολογία των JavaServer PagesTM (JSP), προσφέρει ένα απλοποιημένο, γρήγορο τρόπο για να δημιουργήσει ιστοσελίδες για την εμφάνθση υλικού που παράγεται δυναμικά. Η JSP τεχνολογία είχε ως σκοπό να καταστήσει ευκολότερη και ταχύτερη τη δημιουργία web-based εφαρμογών που λειτουργούν με μια ευρεία ποικιλία από web servers, διακομιστές εφαρμογών, μηχανές αναζήτησης και εργαλεία ανάπτυξης.

Η τεχνολογία JSP επιταχύνει, κατά κάποιον τρόπο την ανάπτυξη δυναμικών ιστοσελίδων με τους εξής τρόπους:

1. *Διαχωρίζει την παραγωγή περιεχομένου από την παρουσίαση της σελίδας.*
   Χρησιμοποιώντας την τεχνολογία JSP, οι προγραμματιστές ιστοσελίδων κάνουν χρήση html ετικετών για το σχεδιασμό και τη μορφή της σελίδας.Χρησιμοποιούν JSP ετικέτες ή scriptlets για να δημιουργήσουν το δυναμικό περιεχόμενο (το περιεχόμενο που αλλάζει ανάλογα με το αίτημα, όπως το αίτημα για τα στοιχεία του λογαριασμού).Η λογική με την οποία παράγεται το περιεχόμενο, βρίσκεται σε ετικέτες και JavaBeans στοιχεία και συνδέονται όλα μεταξύ τους σε scriptlets, τα οποία εκτελούνται από την πλευρά του server. Αν η λογική του πυρήνα βρίσκεται σε tags και beans, στη συνέχεια, άλλα άτομα, όπως web masters και οι σχεδιαστές σελίδων, μπορούν να επεξεργαστούν τη σελίδα JSP χωρίς να επηρεάζουν την παραγωγή του περιεχομένου.

Στο διακομιστή, μια μηχανή JSP ερμηνεύει JSP tags και scriptlets, παράγει περιεχόμενο (για παράδειγμα, με την πρόσβαση σε JavaBeans στοιχεία, την πρόσβαση σε μια βάση δεδομένων με την τεχνολογία JDBCTM ,ή περιλαμβάνοντας αρχεία), και στέλνει τα αποτελέσματα πίσω σε μορφή HTML (ή XML) σελίδας στον browser. Αυτό βοηθά στην προστασία της βιομηχανικής ιδιοκτησίας συγγραφείς κώδικα ,εξασφαλίζοντας παράλληλα πλήρη φορητότητα για οποιοδήποτε web browser, που βασίζεται σε HTML.

2. *Δίνοντας έμφαση σε επαναχρησιμοποιήσιμα στοιχεία.*
   Οι περισσότερες JSP σελίδες βασίζονται σε επαναχρησιμοποιήσιμα, cross- platform στοιχεία (JavaBeans ή Enterprise JavaBeans στοιχεία), για να εκτελέσουν τις περισσότερο πολύπλοκες επεξεργασίες που απατούνται από την εφαρμογή.Οι προγραμματιστές μπορούν να μοιράζονται στοιχεία τα οποία εκτελούν κοινές λειτουργίες ,ή να τα καθιστούν διαθέσιμα σε πιο ευρείες κοινότητες χρηστών και πελατών.Η προσέγγιση που βασίζεται στα στοιχεία (components) ,επιταχύνει πάνω από όλα την ανάπτυξη και επιτρέπει στους οργανισμούς να μοχλεύουν την υπάρχουσα τεχνογνωσία και να αναπτύσοουν προσπάθειες για βέλτιστα αποτελέσματα.

3. *Απλοποιεί την ανάπτυξη των σελίδων με τη χρήση ετικετών (tags)*
   Οι προγραμματιστές web δεν έχουν πάντα την οικειότητα με τις scripting γλώσσες προγραμματισμού.Η τεχνολογία JSP ενσωματώνει πολλή από την λειτουργικότητα που απαιτείται για τη δημιουργία δυναμικού περιεχομένου ,με έναν πολύ εύκολο στη χρήση τρόπο, τις JSP- ειδικές XML ετικέτες.Οι JSP ετικέτες μπορούν να έχουν πρόσβαση και να δημιουργούν στιγμιότυπα JavaBeans στοιχείων , να καθορίζουν ή να ανακτούν bean χαρακτηριστικά, να «κατεβάζουν» applets και να εκτελούν και άλλες λειτουργίες οι οποίες είναι περισσότερο δύσκολες και χρονοβόρες ,να αναπτυχθεί ο κώδικας τους.

Η JSP τεχνολογία μπορεί να επεκταθεί ,μέσω της ανάπτυξης των προσαρμοσμένων βιβλιοθηκών των ετικετών.Σιγά – σιγά ,προγραμματιστές και άλλοι θα μπορέσουν να δημιουργήσουν τις δικές του βιβλιοθήκες ετικετών για κοινές λειτουργίες. Αυτό δίνει τη δυνατότητα στους προγραμματιστές web να εργαστούν με γνώριμα εργαλεία και μεθόδουν κατασκευής, όπως είναι οι ετικέτες, για να εκτελέσουν πολύπλοκες λειτουργίες.

Η τεχνολογία JSP ενσωμάτωνεται εύκολα σε μια ποικιλία αρχιτεκτονικών εφαρμογής, αξιοποιώντας τα υπάρχοντα εργαλεία, και κλιμακώνοντας της θέση της για την υποστήριξη κατανεμημένων εφαρμογών σε όλες τις επιχειρήσεις.Ως μέρος της οικογένειας της τεχνολογίας Java ,και ως αναπόσπαστο κομμάτι της Java 2 ,αρχιτεκτονικής Enterprise Edition ,η τεχνολογία JSP μπορεί να υποστηρίξει εξαιρετικά πολύπλοκες web εφαρμογές:

- Επειδή η μητρική scripting γλώσσα για τις JSP σελίδες βασίζεται στην Java , και επειδή όλες οι σελίδες JSP μεταγλωττίζονται μέσα σε Java Servlets ,η τεχνολογία JSP έχει όλα τα πλεονεκτήματα της τεχνολογίας Java , συμπεριλαμβάνοντας διαχείριση της μνήμης και της ασφάλειας.

Ως κομμάτι της πλατφόρμας Java , η JSP μοιράζεται τα χαρακτηριστικά της γλώσσας προγραμματισμού Java, που είναι: γράφεις τον κώδικα μια φορά και τον τρέχει οπουδήποτε. Καθώς όλο και περισσότεροι κατσκευαστές προσθέτουν υποστήρθξη JSP στα προϊόντα τους, μπορεί ο καθένας να χρησιμοποιήσει servers και εργαλεία της επιλογής του , να αλλάξει servers και εργαλεία χωρίς να επηρεαστούν καθόλου οι τρέχουσες εφαρμογές.

Πώς φαίνεται μια σελίδα JSP;

Μια σελίδα JSP φαίνεται όπως μια κοινή HTML ή XML σελίδα , με πρόσθετα στοιχεία τα οποία επεξεργάζεται η μηχανή JSP και τα αναπαριστά.Τυπικά, τα στοιχεία JSP δημιουργούν κείμενο το οποίο εισάγεται στη σελίδα με τα αποτελέσματα.

2.2.4 Στατιστικό πακέτο

Το στατιστικό πακέτο το οποίο χρησιμοποιήθηκε στην εφαρμογή για την εύρεση των p-values για κάθε nmer είναι το στατιστικό πρόγραμμα R.

Η γλώσσα και το περιβάλλον R χρησιμοποιείται για στατιστικούς υπολογισμούς και γραφικά. Αποτελεί ένα project του GNU το οποίο είναι παρόμοιο με το περιβάλλον και τη γλώσσα S και ανπτύχθηκε στα εργαστήρια της Bell (formerly AT&T, now Lucent Technologies) από τους John Chambers και τους συναδέλφους του.

Το R παρέχει μια ποικιλία στατιστικών (γραμμικά και μη-γραμμικά μοντέλα, κλασσικά στατιστικά τεστ, ανάλυση χρονικών σειρών, κλπ) και γραφικών τεχνικών και είναι αρκετά επεκτάσιμο.

Ένα από τα ισχυρα πλεονέκτηματα του R είναι η ευκολια με την οποία καλοσχεδιασμένες γραφικές παραστάσεις μπορούν να παραχθούν ,συμπεριλαμβάνοντας μαθηματικά σύμβολα και φόρμουλες όπου είναι ανγκαίο.Έχει ληφθεί ιδιαίτερη φροντίδα για τις default επιλογές στα γραφικά, αλλά ο χρήστης έχει τον απόλυτο έλεγχο.

Το R είναι διαθέσιμο ως ελεύθερο λογισμικό υπό τους όρους του Free Software Foundations του GNU General Public License, σε μορφή source code. Μεταγλωττίζει και εκτελείται σε μια μεγάλη ποικιλία UNIX πλατφορμών και παρόμοιων συστημάτων ,στα Windows στα λειτουργικά συστήματα Mac.

Το περιβάλλον του R

Το R είναι ένα ολοκληρωμένο περιβάλλον λογισμικών δραστηριοτήτων για εκμετάλλευση δεδομένων ,υπολογισμούς και εμφάνιση γραφικών.Περιλαμβάνει:

- Έναν αποτελεσματικό χερισμό δεδομένων και δυνατότητα αποθήκευσης
- Ένα σύνολο χειριστών για υπολογισμούς σε συστοιχίες ,σε συγκεκριμένους πίνακες
- Μια μεγάλη συλλεκτική ολοκληρωμένη συλλογή ενδιάμεσων εργαλείων για ανάλυση δεδομένων
- Γραφικές δυνατότητες και εμφάνιση είτε στην οθόνη είτε σε hardcopy
- Και μία καλά αναεπτυγμένη απλή και αποτελεσματική γλώσσα προγραμματισμού που περιλαμβάνει συνθήκες , βρόχους ,ορισμένες αναδρομικές συναρτήσεις από χρήστες και δυαντότητες εισόδου –εξόδου.

Ο όρος «περιβάλλον» τείνει να να το χαρακτηρίσει ως ένα πλήρως σχεδιασμένο σύστημα ,παρά μια αυξητική επικάθηση πολύ ιδιαίτερων και μη ευέλικτων εργαλείων, όπως είναι συχνά η περίπτωση άλλων λογισμικών ανάλυσης δεδομένων.

Η R ,όπως και η S ,σχεδιάστηκε γύρω από μία αληθινή γλώσσα υπολογιστή και επιτρέπει στους χρήστες να προσθέτουν λειτουργικότητα ορίζοντας νέες συναρτήσεις. Μεγάλο μέρος του συστήματος είναι το ίδιο γραμμένο στη διάλεκτο R της S ,το οποίο κάνει εύκολο στους χρήστες να ακολουθήσουν τις αλγοριθμικές επιλογές που γίνονται. Για ιδιαίτερα υπολογιστικά καθήκοντα η C ,C++ και Fortran κώδικας μπορεί να συνδεθεί και να εκτελεστεί από το περιβάλλον. Προχωρημένοι χρήστες μπορούν να γράψουν κώδικα για να διαχειρίζονται άμεσα R αντικείμενα.

Πολλοί χρήστες θεωρούν το R ώς στατιστικό σύστημα.Το R όμως μπορεί να επεκταθεί εύκολα μέσω πακέτων.Υπάρχουν περίπου οχτώ πακέτα τα οποία παρέχονται με το R και πολύ περισσότερα είναι διαθέσιμα μέσω της οικογένειας CRAN σε sites του internet και καλύπτουν μια μεγάλη ποικιλία σύγχρονης στατιστικής.

Το R έχει το δικό του LaTeX-like documentation format ,το οποίο χρησιμοποιείται για να καλύψει ολοκληρωμένη τεκμηρίωση τόσο σε online διάφορες μορφές όσο και σε έντυπη μορφή.

Χρήση του wilcoxon test

Το wilcoxon-signed rank test είναι ένα μη παραμετρικό στατιστικό υποθετικό test για την περίπτωση δύο σχετιζόμενων δειγμάτων ή επαναλαμβανόμενων μετρήσεων σε ένα ενιαίο δείγμα. Μπορεί να χρησιμοποιηθεί να χρησιμοποιηθεί ως εναλλακτικό στο στο t-test όταν ο πληθυσμός δεν μπορεί να θεωρηθεί ότι ακολουθεί την κανονική κατανομή.

Το wilcoxon test περιλαμβάνει συγκρίσεις διαφορών μεταξύ μετρήσεων ,γι’αυτό απαιτεί ότι τα δεδομένα μετρούνται σε ένα επίπεδο διαστήματος της μέτρησης. Ωστόσο δεν απαιτεί υποθέσεις για τη μορφή της κατανομής των μετρήσεων.

Με τη χρήση του στατιστικού τεστ βρίσκουμε τις pvalues για κάθε nmer.Μια pvalue είναι η πιθανότητα να αποκτήσουμε στατιστικό αποτέλεσμα της δοκιμής ,τουλάχιστον τόσο μεγάλο όσο αυτό που παρατηρήθηκε , υποθέτοντας ότι η μηδενική υπόθεση είναι αληθής.

Όσο μικρότερη είναι η τιμή της pvalue , τόσο λιγότερο πιθανό είναι το αποτέλεσμα αν η αρχική υπόθεση είναι αληθής ,και συνεπώς τόσο πιο σημαντικό το αποτέλεσμα στην ευρύτερη έννοια της στατιστικής σημασίας.

2.2.5 Υπόλοιπα στοιχεία υλοποίησης

Ο server που χρησιμοποιήθηκε για την εφαρμογή είναι ο Apache Tomcat Server 6.0 και χρησιμοποιήθηκε ενσωματωμένος με το ολοκληρωμένο περιβάλλον ανάπτυξης της εφαρμογής NetBeans 6.8 IDE.

Ως βάση δεδομένων χρησιμοποιήθηκε η γνωστή MySQL και συγκεκριμένα ο MySQL Server 5.1,και περιέχει τη βάση δεδομένων bio και τον πίνακα data.

Για τις στατικές σελίδες έχει χρησιμοποιηθεί η γλώσσα html.

Γενικότερα η εφαρμογή είναι δυναμική και εφόσον σε αρχικό επίπεδο έχει πραγματοποιηθεί η λειτουργία της ,σε δεύτερη φάση μπορούν να προστεθούν στοιχεία τα οποία θα δίνουν πιο εντυπωσιακή εμφάνιση ,πληροφορίες για τους κατασκευαστές του περιβάλλοντος και επικοινωνία μαζί τους.

Ο υπολογιστής στον οποίο έτρεξε η εφαρμογή έχει λειτουργικό σύστημα Windows 7.

Ο επεξεργαστής είναι Intel® Core™ 2 Duo με ταχύτητα στα 2.40 GHz.

Η μνήμη του υπολογιστή είναι στα 3.00 GB.

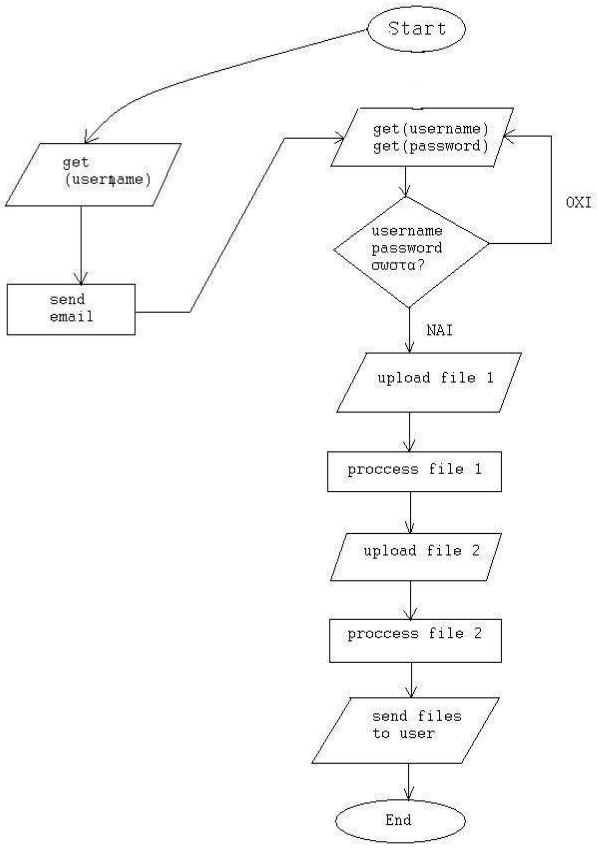
ΚΕΦΑΛΑΙΟ 3

Αποτελέσματα

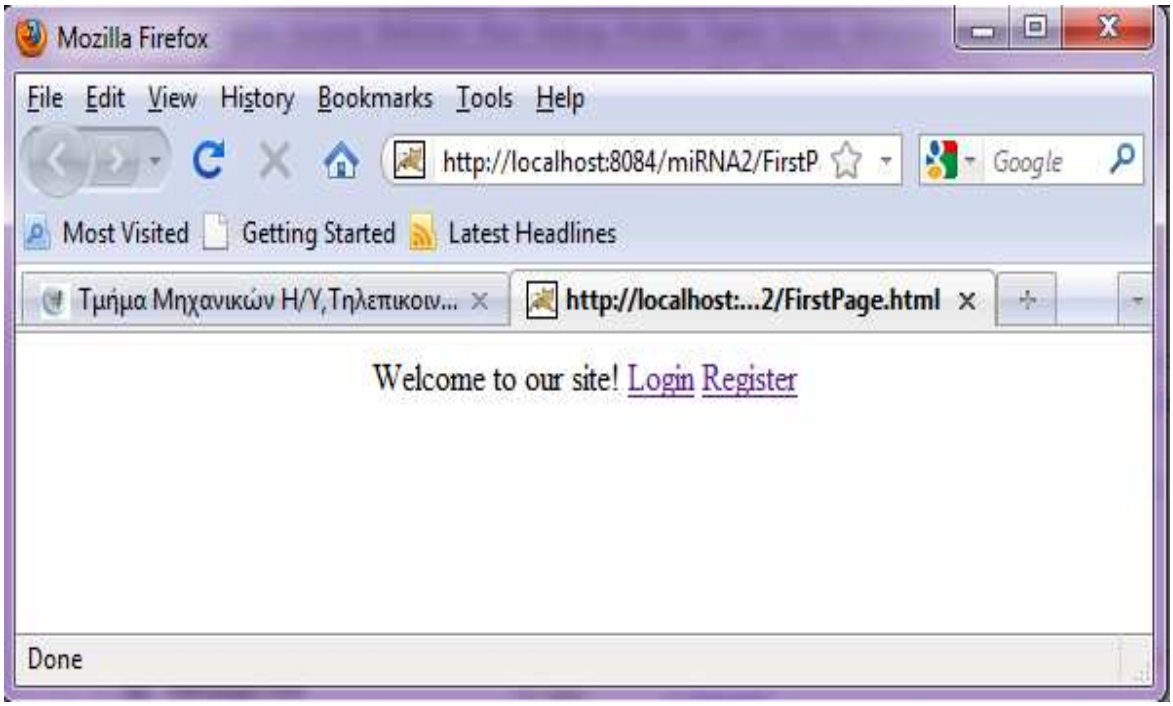
3.1 Flowchart εφαρμογής

Συνολικά η εφαρμογή αποτελείται από 21 αρχεία.Τα 7 από τα αρχεία αυτά είναι αρχεία jsp και αφορούν τις δυναμικές σελίδες, τα 2 είναι αρχεία html και αφορούν τις στατικές σελίδες ,τα 11 είναι αρχεία java και αφορούν την κύρια επεξεργασία των δεδομένων και το τελευταίο αρχείο είναι το αρχείο R για τη στατιστική επεξεργασία.

Παρακάτω φαίνεται το διάγραμμα ροής(flowchart) της εφαρμογής ,πως δηλαδή συνδέονται μεταξύ τους τα επιμέρους αρχεία ,οι συνθήκες ,η είσοδος των αρχείων και η έξοδος.

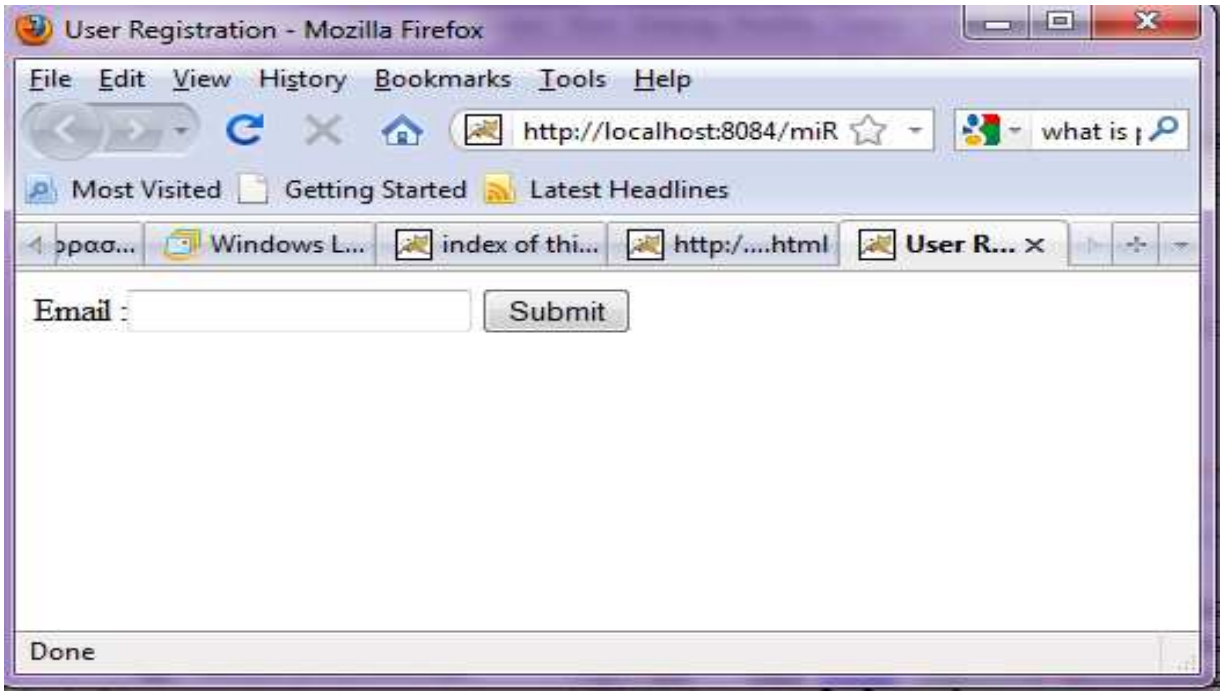


3.2 Παράδειγμα χρήσης της εφαρμογής

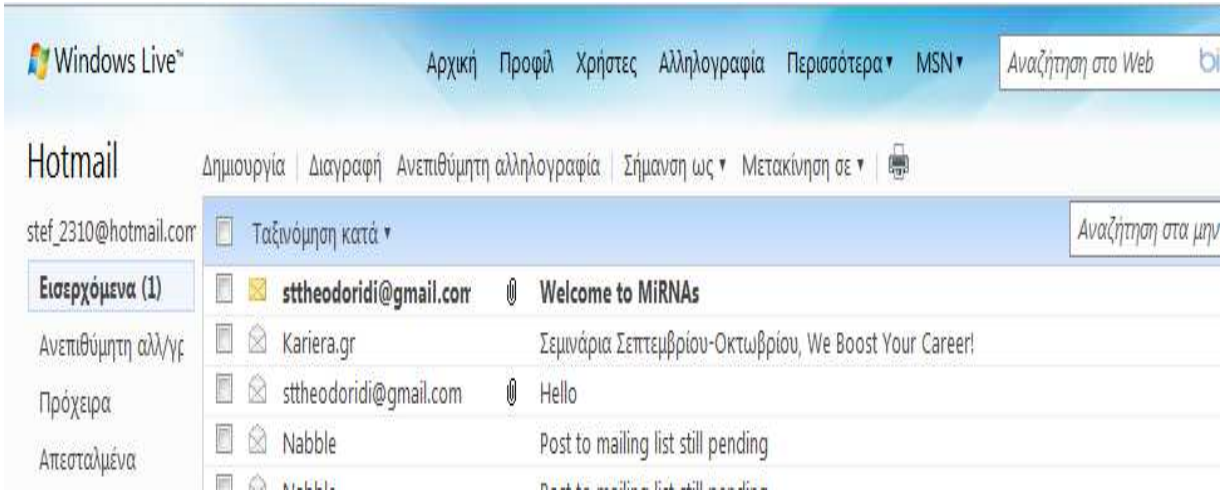


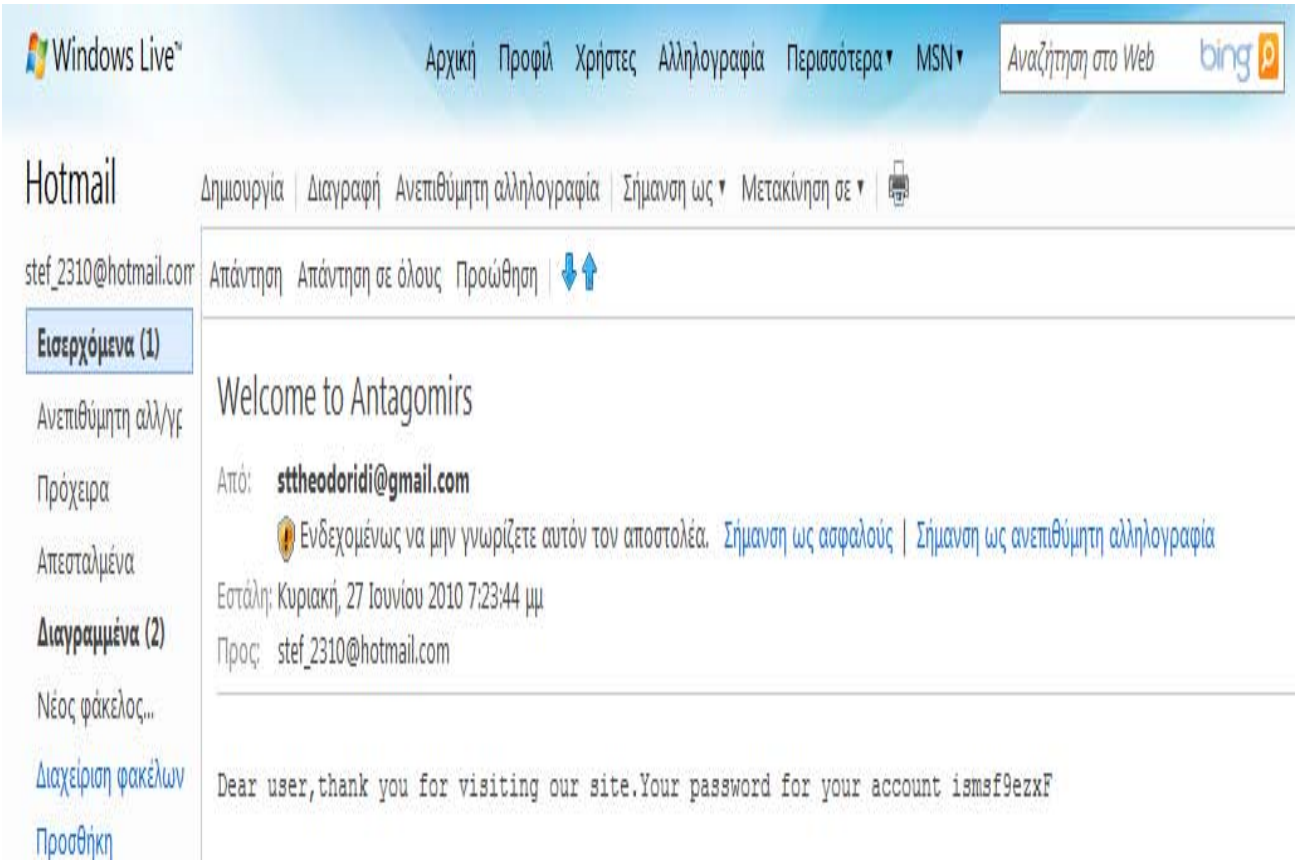
Παραπάνω φαίνεται η αρχική σελίδα της εφαρμογής ,όπου ο χρήστης επιλέγει σε ποια σελίδα θέλει να μεταφερθεί. Αν είναι ήδη εγγεγραμμένος χρήστης τότε μπορεί να μεταφερθεί αμέσως στη σελίδα για να κάνει εισαγωγή.Διαφορετικά θα πρέπει πρώτα να κάνει την εγγραφή και στη συνέχεια να μεταφερθεί στη σελίδα εισαγωγής.

Η αρχική σελίδα της διεπαφής φαίνεται παρακάτω. Είναι η σελίδα όπου ο χρήστης κάνει εγγραφή στο σύστημα δίνοντας το email ως username,εφόσον το email για κάθε χρήστη είναι μοναδικό, και στη συνέχεια λαμβάνει ένα mail με τον κωδικό του.



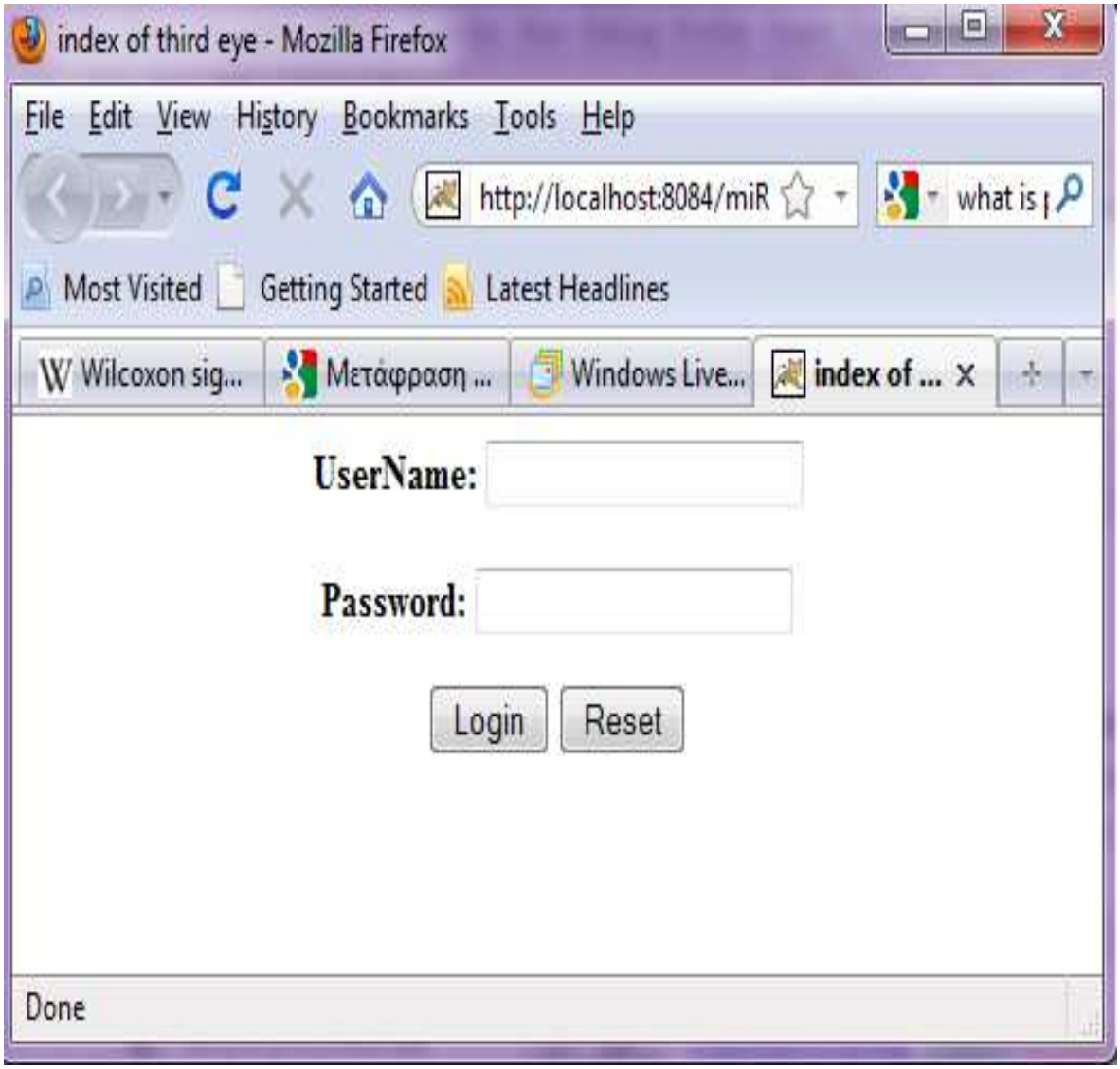
Παρακάτω ακολουθούν οι σελίδες όπου φαίνεται το email που λαμβάνει ο χρήστης και περιέχει τον κωδικό ,ο οποίος όπως είπαμε έχει δημιουργηθεί από τυχαία γεννήτρια αλαφριθμητικών.



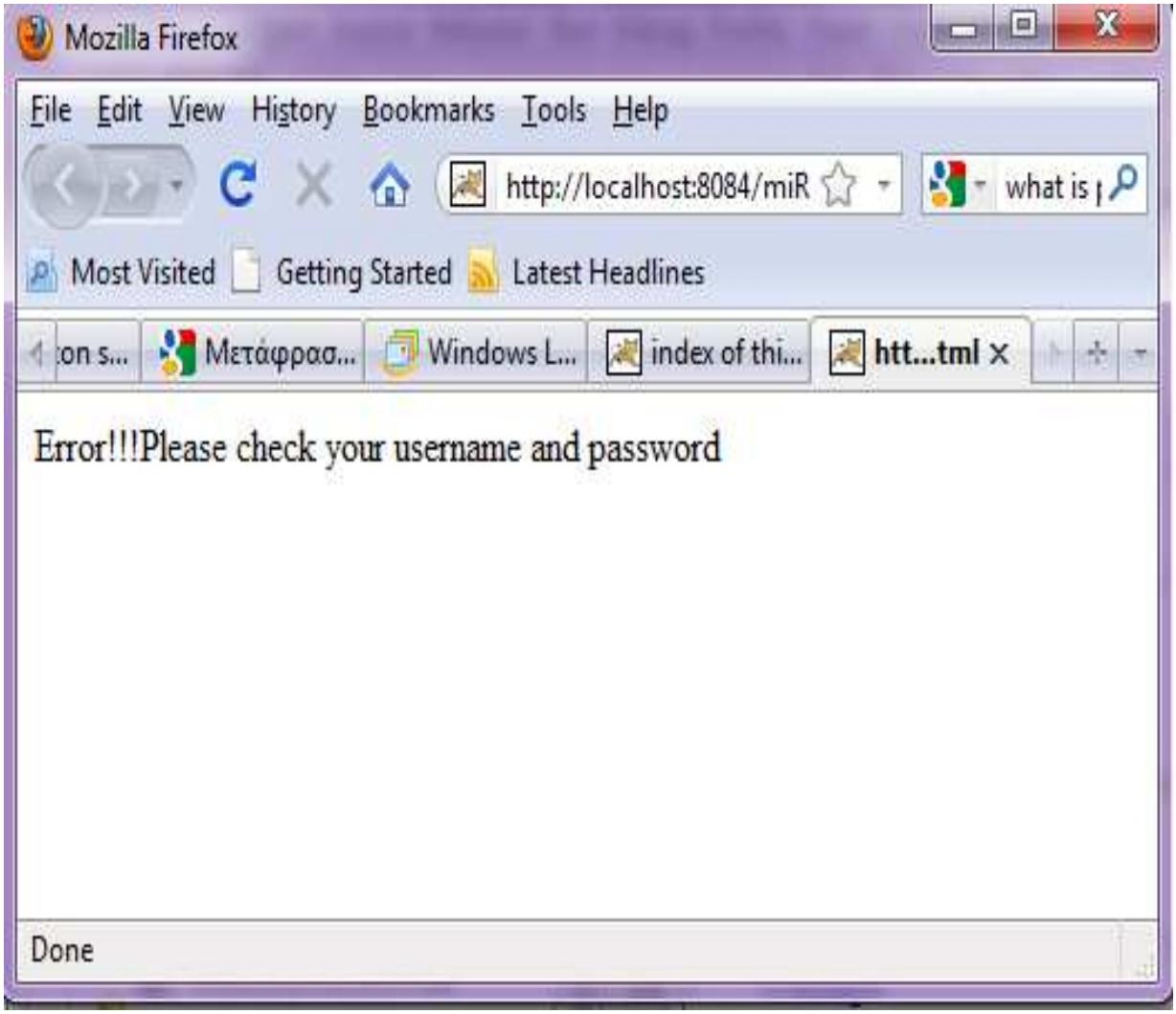


Παρακάτω φαίνεται αρχικά η σελίδα στην οποία μεταφέρεται ο χρήστης μετά την εγγραφή του στο σύστημα και είναι η σελίδα εισαγωγής στο σύστημα. Γίνεται επικοινωνία με τη βάση δεδομένων και εφόσον ο χρήστης έχει δώσει τα σωστά στοιχεία τότε μεταφέρεται στη σελίδα όπου ανεβάζει τα αρχεία ,διαφορετικά μεταφέρεται στη σελίδα που τον ενημερώνει να ελέγξει τα στοιχεία του.

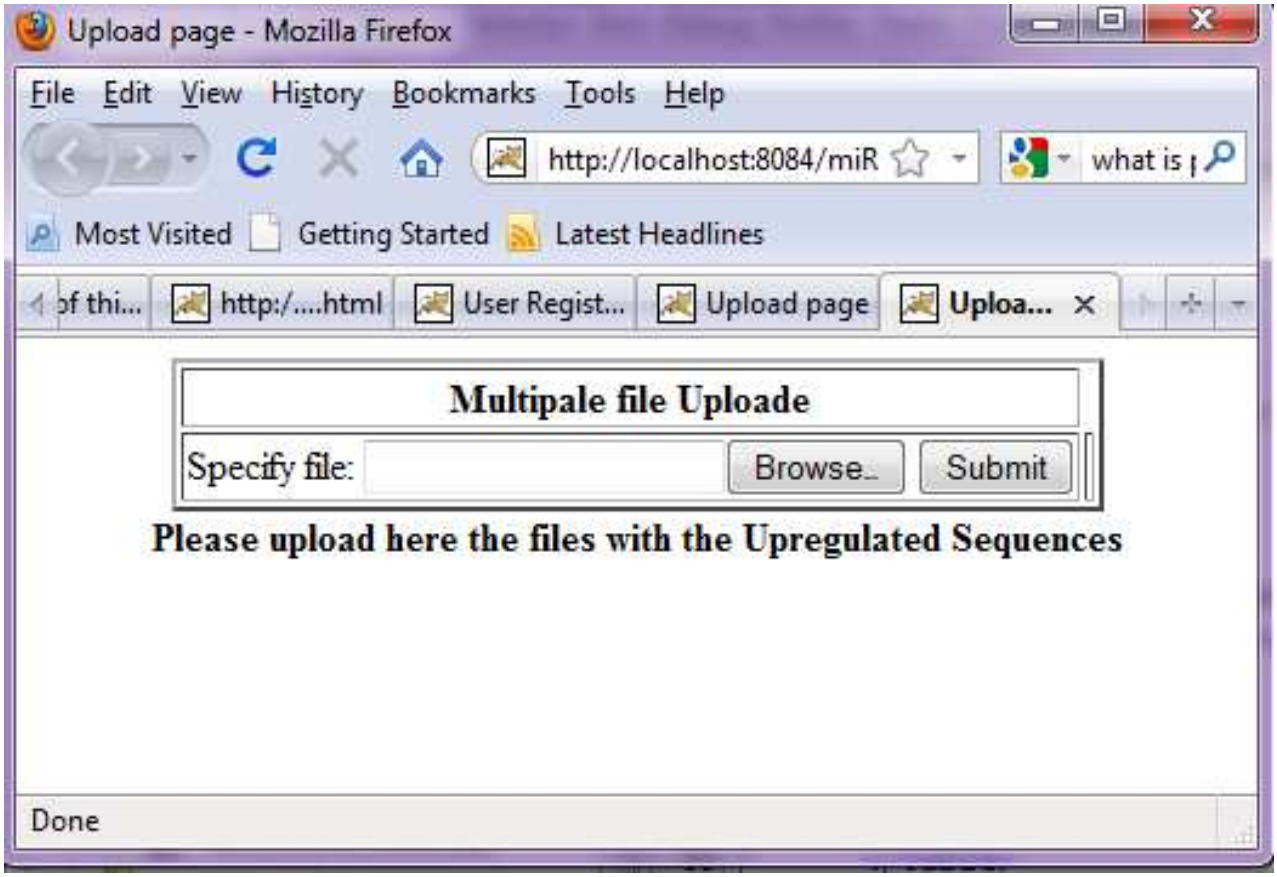
Είναι αντίστοιχα οι σελίδες loginpage.jsp και error.html.



Σελίδα error.html.Ενημέρωση του χρήστη για λάθος στοιχεία.

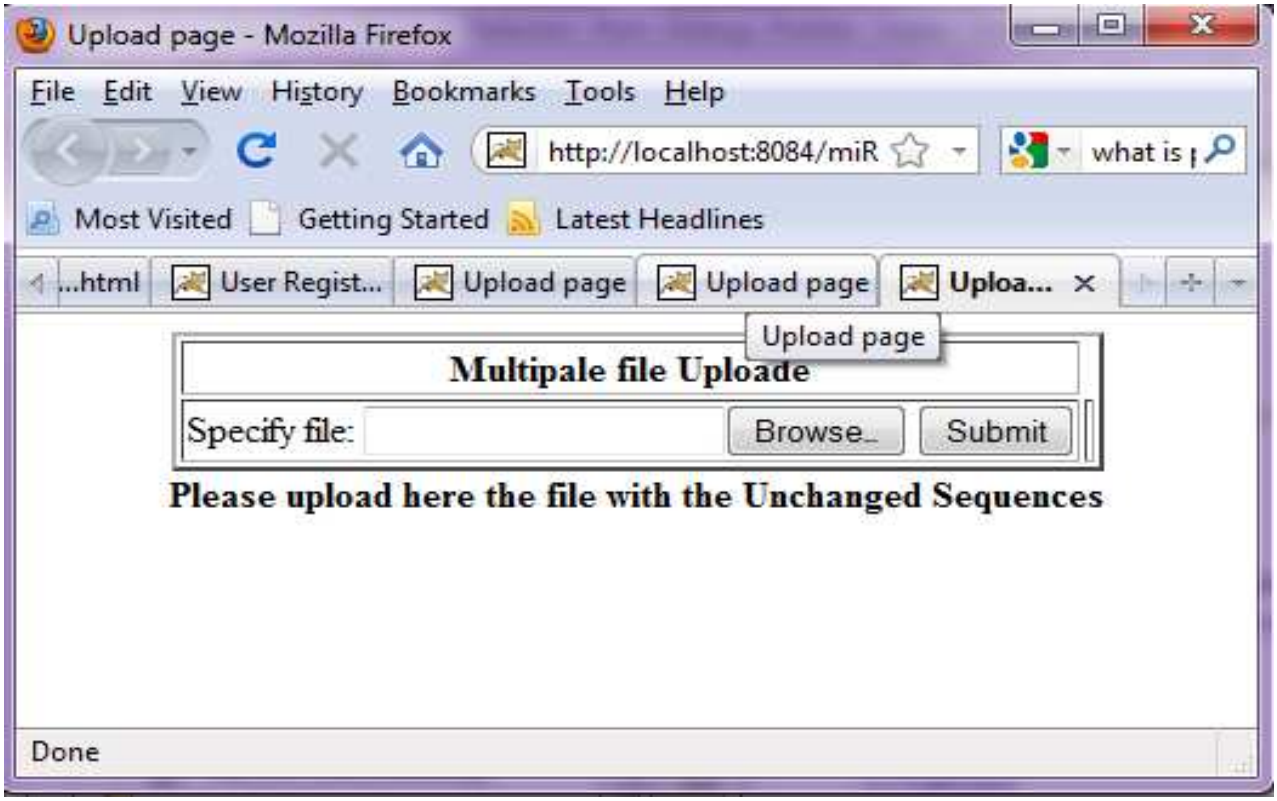


Σελίδα file_multipale.jsp.Ο χρήστης έχει δώσει σωστά τα στοιχεία του και είναι έτοιμος για το ανέβασμα των αρχείων στον server για επεξεργασία. Στη συγκεκριμένη φόρμα ανβάζει το αρχείο που περιέχει τις ακολουθίες UP. Αυτό γίνεται για να αποφύγουμε το λάθος ανέβασμα αρχείων και συνεπώς αναντίστοιχων αποτελεσμάτων.



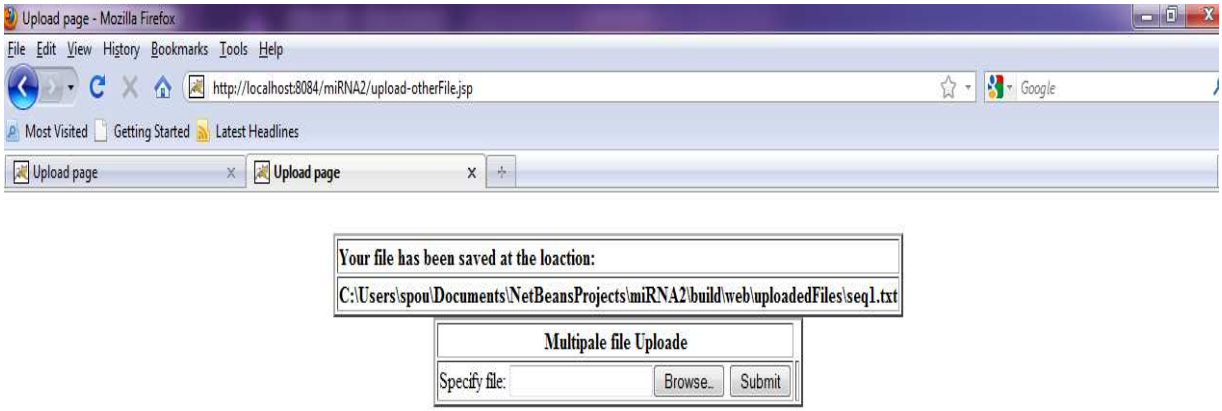
Αφού ο χρήστης ανεβάσει το πρώτο αρχείο στη συνέχεια μεταφέρεται στη σελίδα για το ανέβασμα και του δεύτερου αρχείου, ενώ η επεξεργασία του πρώτου αρχείου έχει ξεκινήσει.

Παρακάτω φαίνεται η σελίδα όπου ζητάει στον χρήστη να ανεβάσει και το αρχείο με τις ακολουθίες Unchnged.

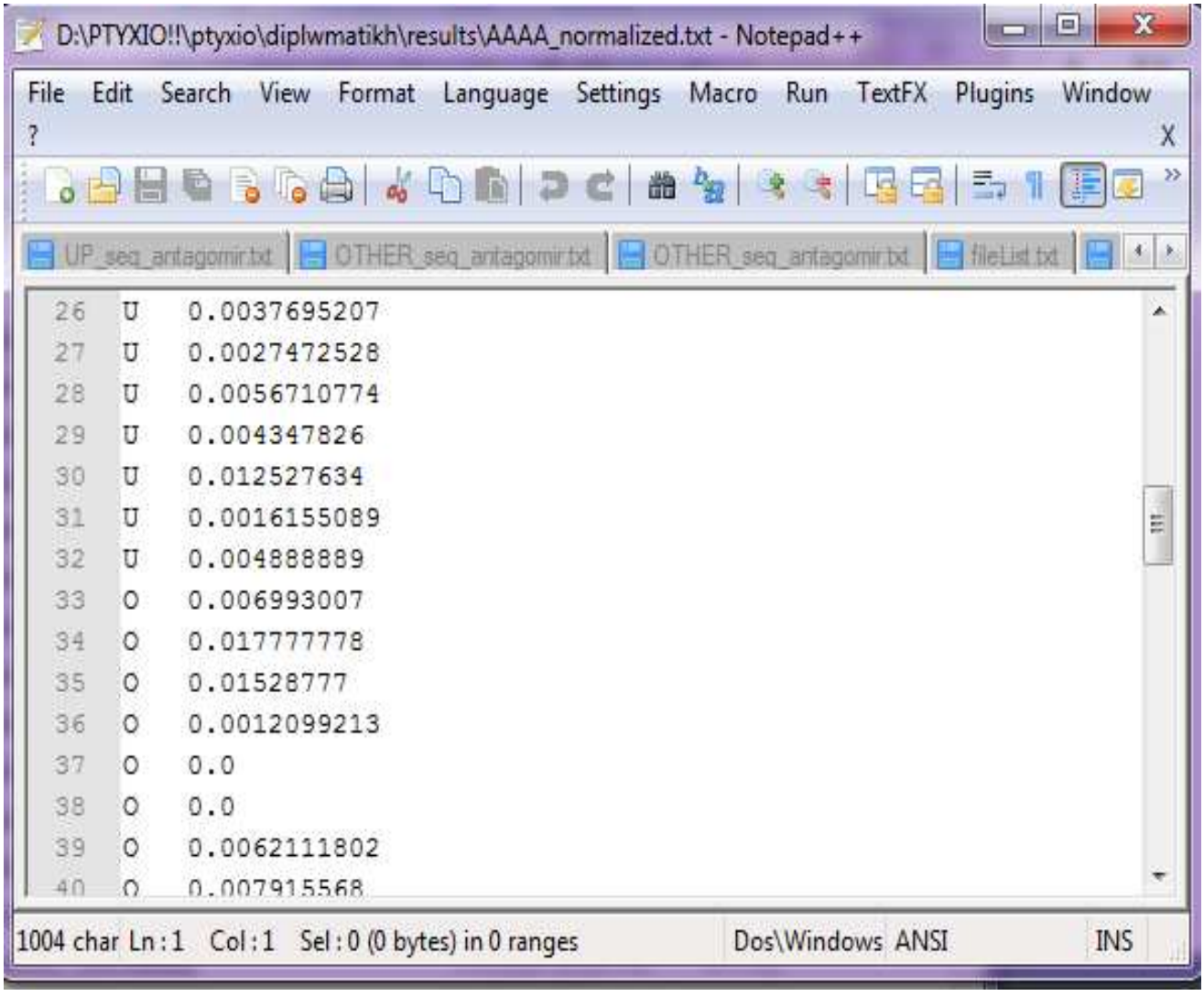


Εφόσον ο χρήστης έχει ανεβάσει και το δεύτερο αρχείο του ξεκινάει και η επεξεργασία του.

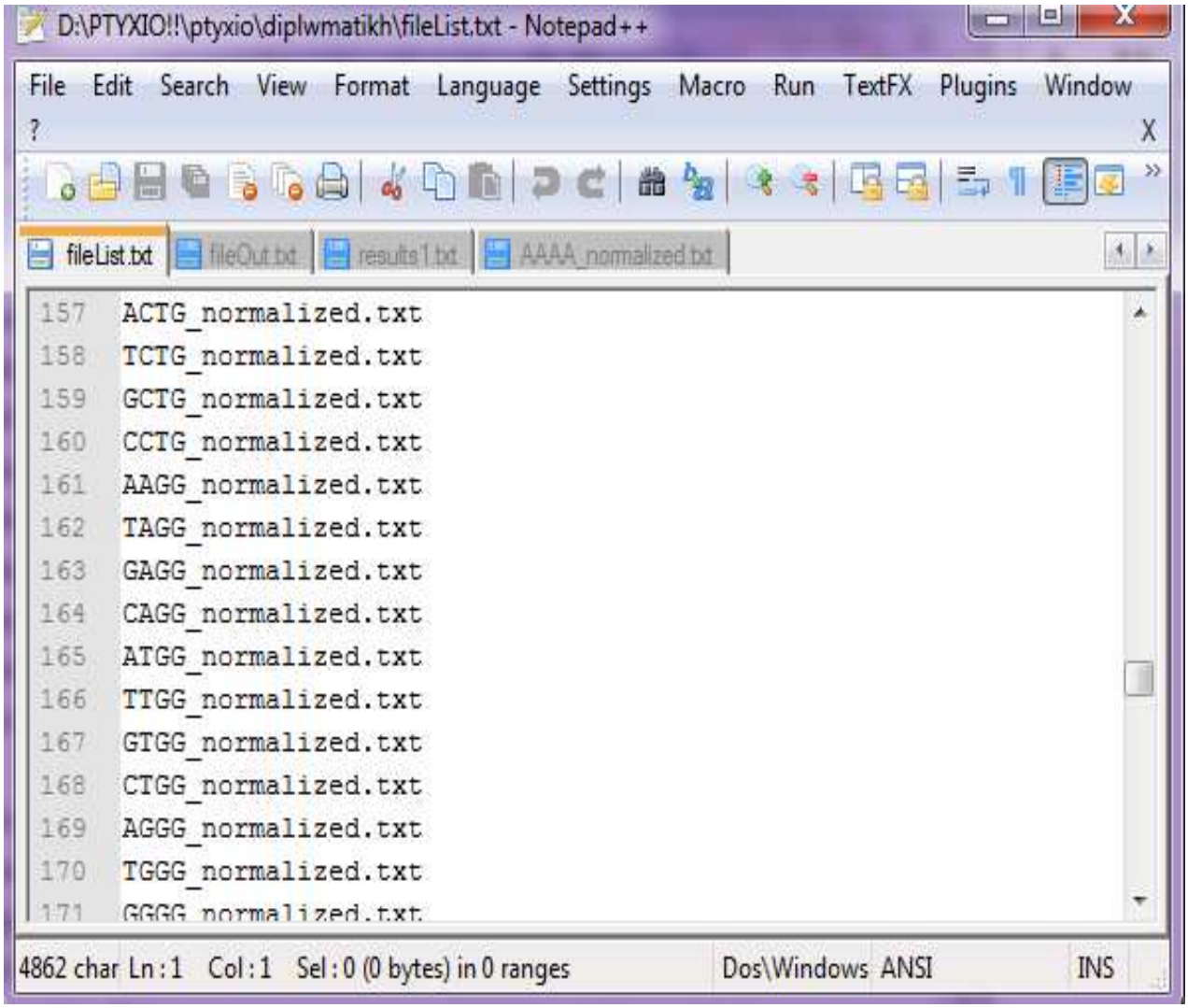
Ο χρήστης μεταφέρεται στη σελίδα όπου τον ενημερώνει ότι σε λίγη ώρα θα λάβει τα αποτελέσματα.

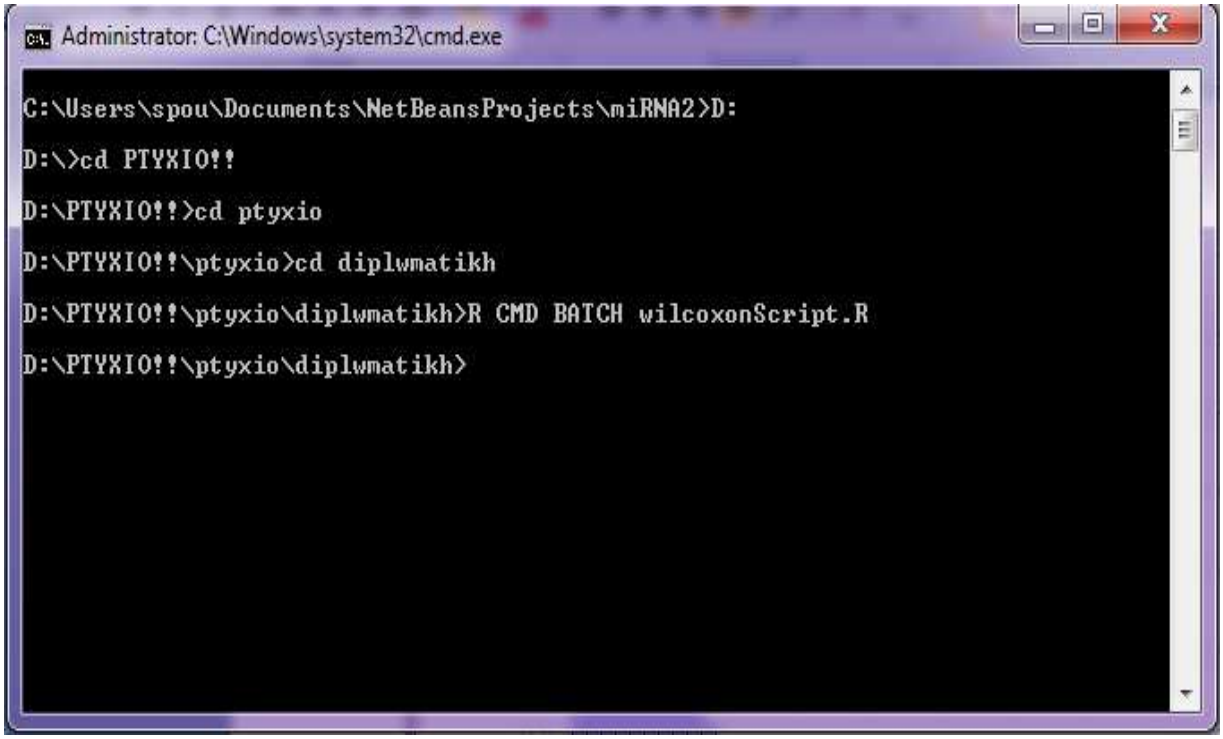


Στις σελίδες που ακολουθούν φαίνονται τα αποτελέσματα της επεξεργασίας των αρχείων του χρήστη. Αρχικά δημιουργούνται για κάθε ένα nmer ένα αρχείο το οποίο μέσα περιέχει το σύμβολο της ακολουθίας στην οποιά ανήκει U για τις Upregulated και O για τις Unchnged. Δίπλα από κάθε σύμβολο εμφανίζεται και το αποτέλεσμα του αλγορίθμου που βρίσκει όλες τις non-overlapping εμφανίσεις του εκάστοτε nmer σε κάθε ακολουθία όπως αυτές διαβάζονται από τα αντίστοιχα αρχεία και τα αποτελέσματα εμφανίζονται για κάθε ακολουθία με τη σειρά που διαβάστηκε.

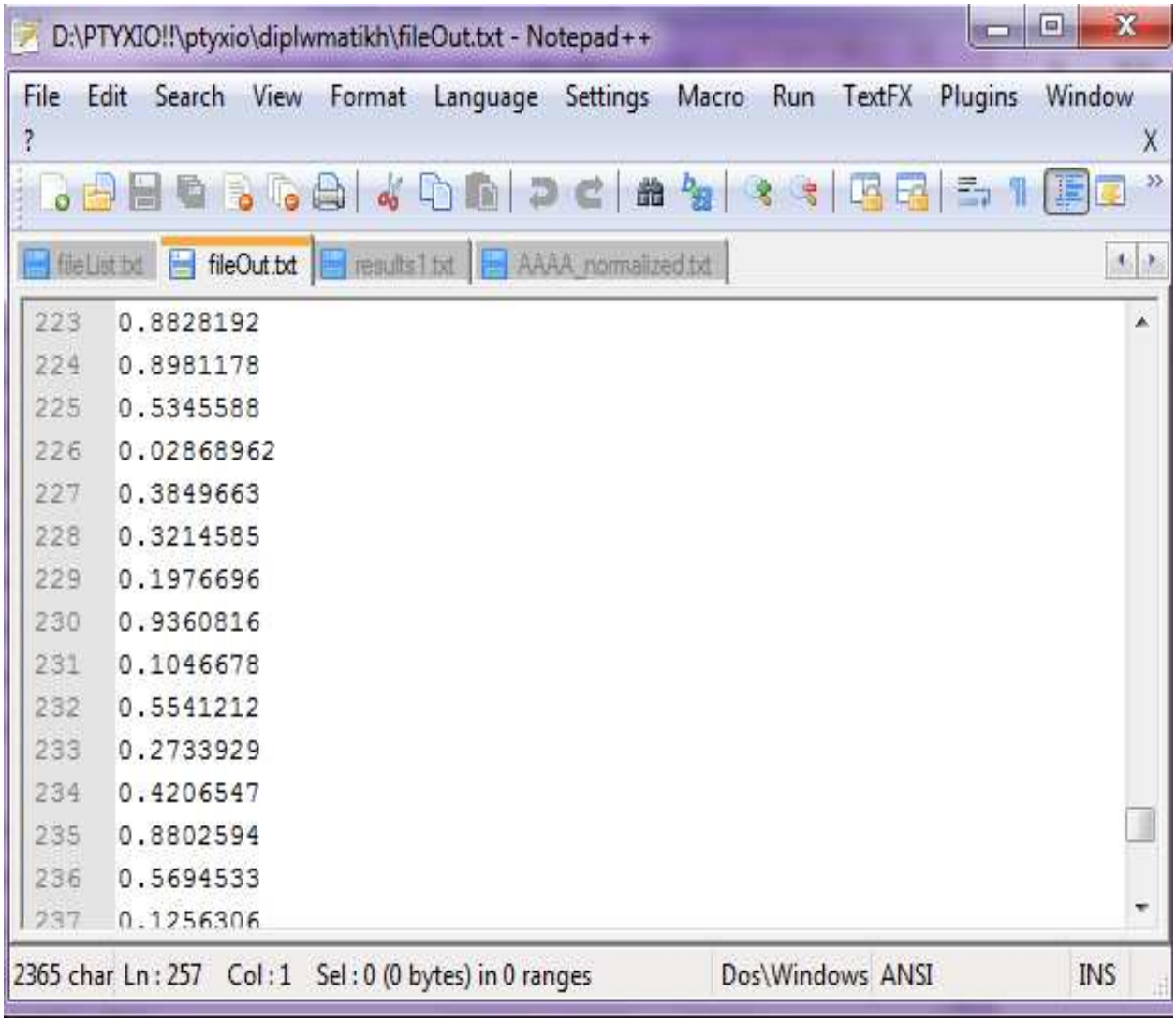


Το δεύτερο αρχείο που δημιουργείται μετά την πρώτη ανάγνωση των αρχείων είναι το αρχείο fileList.txt και φαίνεται παρακάτω.Στο αρχείο fileList.txt εμφανίζονται τα ονόματα των αρχείων που έχουν δημιουργηθεί για κάθε nmer και έχουν με τη σειρά που εμφανίζονται τα nmer στη λίστα ,που την επιστρέφει το πρόγραμμα παραγωγής όλων των πιαθανων ακολουθιων ,μήκους της επιλογής μας από το αλφάβητο {A,T,C,G}.Στην περίπτωση μας το μήκος είναι 4.





Αφού λοιπόν έχουν δημιουργηθεί τα παραπάνω αρχεία στη συνέχεια γίνεται η στατιστική επεξεργασία των δεδομένων.Το πρόγραμμα που καλείται αυτόματα είναι το wilcoxonScript.R και καλείται με την εντολή R CMD BATCH wilcoxonScript.R από το command prompt των Windows όπως φαίνεται παραπάνω, μέσω ενός αρχείου bat και εκτελείται .Το wilcoxonScript.R παίρνει ως εισόδους τα αρχεία fileList και για κάθε ένα nmer το αντίστοιχο αρχείο του με τα normalized counts. Η έξοδος του πργράμματος είναι το αρχείο fileOut.txt στο οποίο αναγράφονται οι τιμές pvalues για κάθε nmer όπως αυτά εμφανίζονται στο αρχείο fileList.txt.

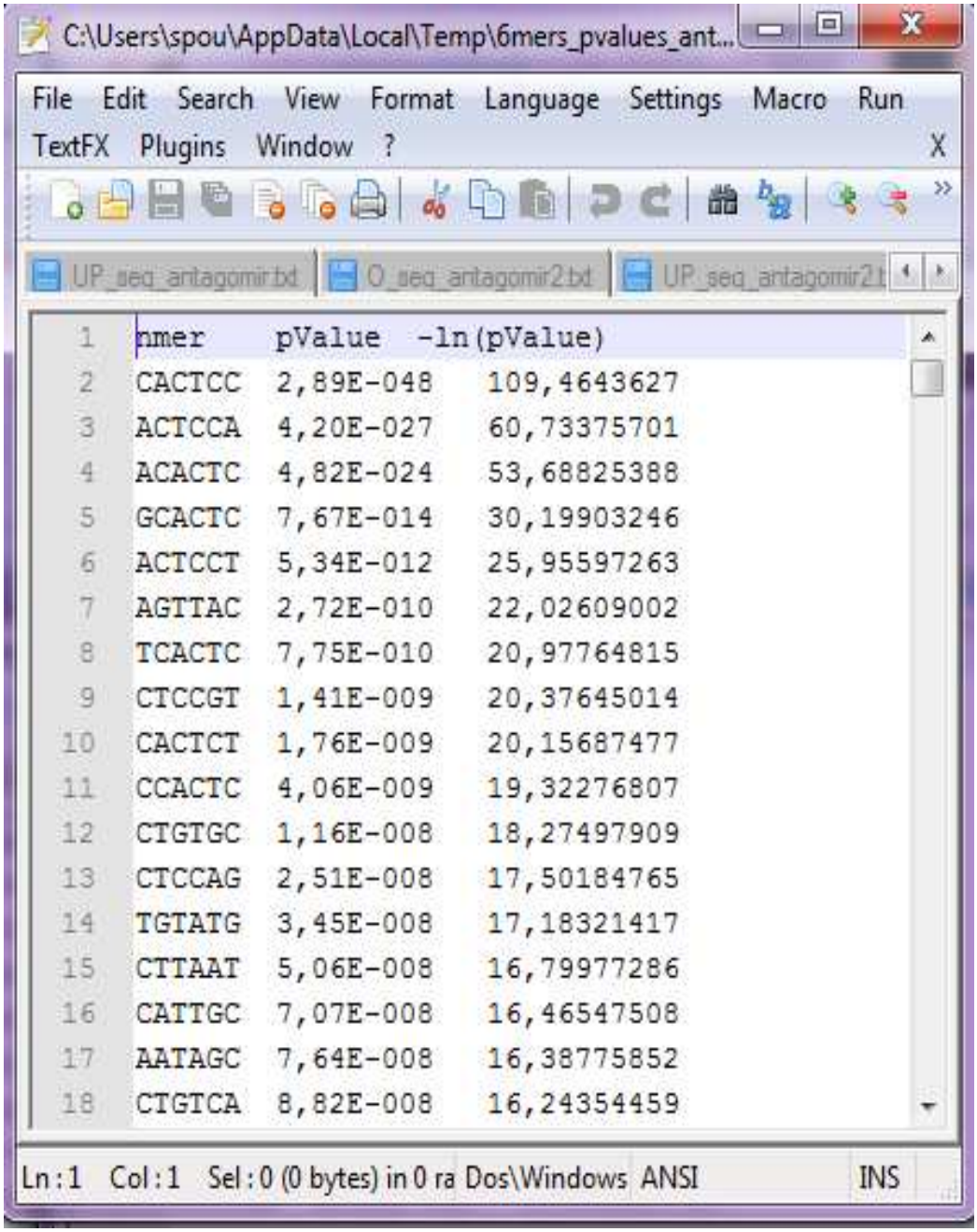


Σε αυτή τη σελίδα φαίνεται το αποτέλεσμα του wilcoxonScript ,που είναι το αρχείο fileOut με τις τιμές των pvalues για κάθε nmer.

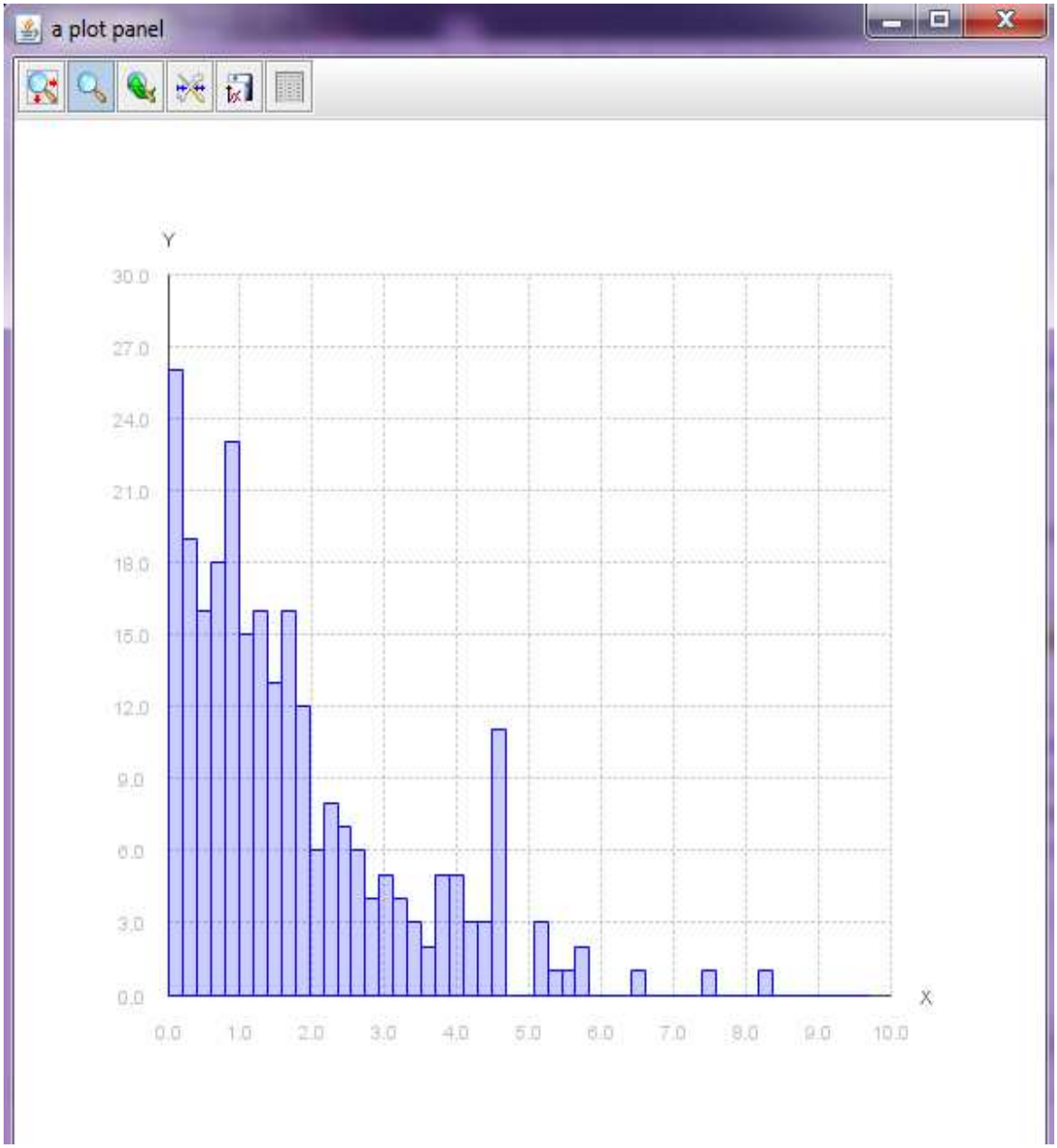
Στη συνέχεια υπολογίζουμε τα –lnpvalues για κάθε pvalue και ταξινομούμε σύμφωνα με τις τιμές των –lnpvalues από τη μεγαλύτερη στη μικρότερη.(descending order)

Στη συνέχεια γράφουμε τα ταξινομημένα αποτελέσματα στο αρχείο results1.txt, όπου περιέχονται το κάθε nmer,η αντίστοιχη pvalue και η αντίστοιχη –lnpavalue.

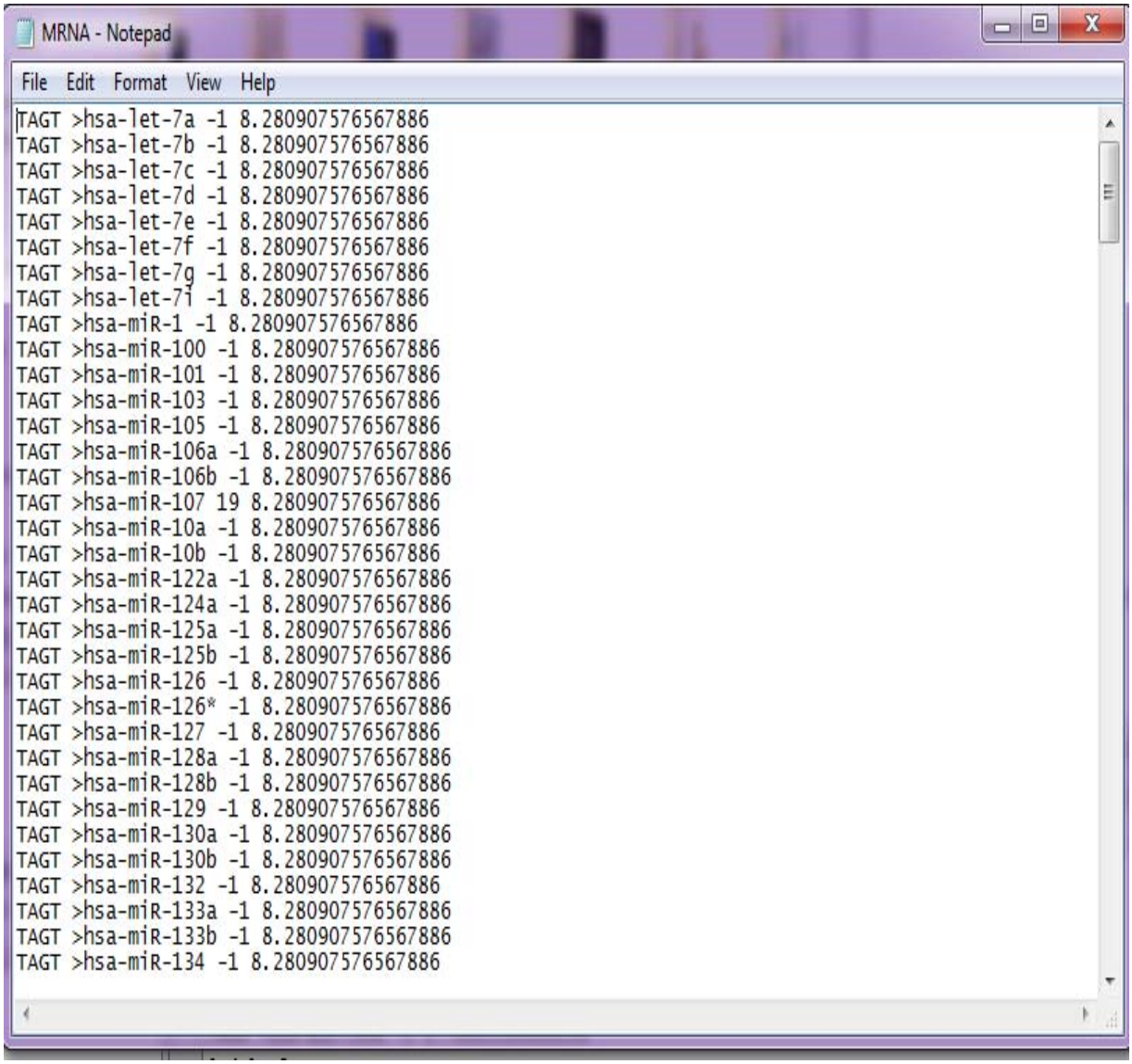
Το αρχείο results1.txt φαίνεται παρακάτω.



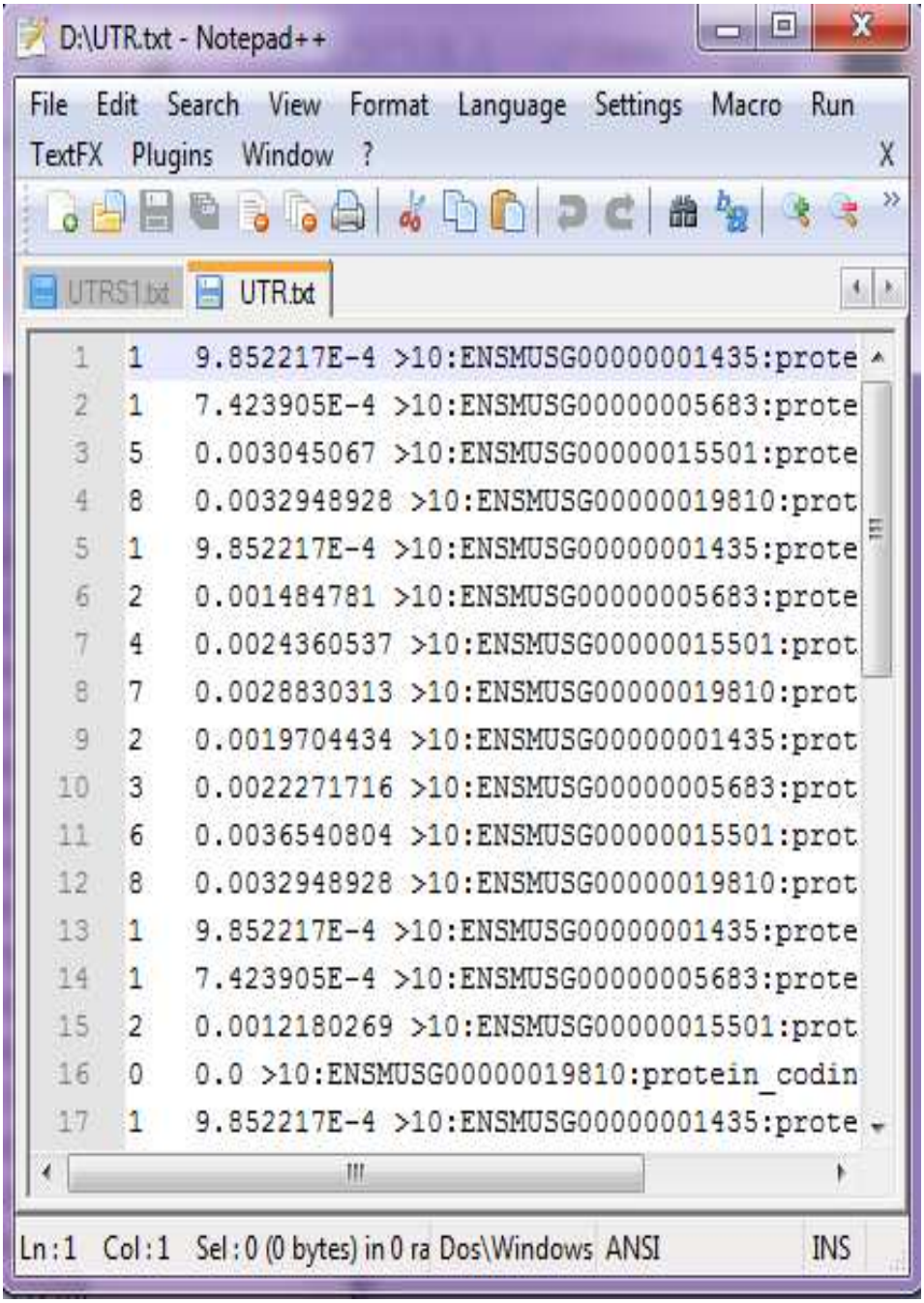
	nmer	pValue	-ln(pValue)
2	CACTCC	2,89E-048	109,4643627
3	ACTCCA	4,20E-027	60,73375701
4	ACACTC	4,82E-024	53,68825388
5	GCACTC	7,67E-014	30,19903246
6	ACTCCT	5,34E-012	25,95597263
7	AGTTAC	2,72E-010	22,02609002
8	TCACTC	7,75E-010	20,97764815
9	CTCCGT	1,41E-009	20,37645014
10	CACTCT	1,76E-009	20,15687477
11	CCACTC	4,06E-009	19,32276807
12	CTGTGC	1,16E-008	18,27497909
13	CTCCAG	2,51E-008	17,50184765
14	TGTATG	3,45E-008	17,18321417
15	CTTAAT	5,06E-008	16,79977286
16	CATTGC	7,07E-008	16,46547508
17	AATAGC	7,64E-008	16,38775852
18	CTGTCA	8,82E-008	16,24354459



Εδώ φαίνεται το ιστόγραμμα των –lnpvalues.



Τέλος για τα πρώτα nmers των οποίων η lnpvalue ξεχωρίζει, προσπαθούμε να τα ανζητήσουμε στις πρώτες θέσεις των MicroRNAs. Το πρόγραμμα λοιπόν διαβάζει το αρχείο το οποίο περιέχει ακολουθίες microRNA σε FASTA μορφή και προσπαθεί να βρει αν στις πρώτες θέσεις του καθενός υπάρχει κάποιο από τα nmers που ξεχωρίζουν σύμφωνα με την lnpvalue τους.Στη συνέχεια επιστρέφει στον χρήστη ένα αρχείο το οποίο περιέχει το όνομα του nmer ,το όνομα του microRNA, τη θέση στην οποία βρίσκεται το nmer και την τιμή lnpvalue. Τέλος για κάθε ένα nmer , από τα παραπάνω που προέκυψαν δημιουργείται ένα αρχείο το οποίο περιέχει τα UTRs στα οποία έχουν βρεθεί καθώς και τη θέση τους σε αυτά.


D:\UTR.txt - Notepad++
File Edit Search View Format Language Settings Macro Run
TextFX Plugins Window ?
UTRS1.txt
UTR.txt
1 1 9.852217E-4 >10:ENSMUSG00000001435:prote
2 1 7.423905E-4 >10:ENSMUSG00000005683:prote
3 5 0.003045067 >10:ENSMUSG00000015501:prote
4 8 0.0032948928 >10:ENSMUSG00000019810:prot
5 1 9.852217E-4 >10:ENSMUSG00000001435:prote
6 2 0.001484781 >10:ENSMUSG00000005683:prote
7 4 0.0024360537 >10:ENSMUSG00000015501:prot
8 7 0.0028830313 >10:ENSMUSG00000019810:prot
9 2 0.0019704434 >10:ENSMUSG00000001435:prot
10 3 0.0022271716 >10:ENSMUSG00000005683:prot
11 6 0.0036540804 >10:ENSMUSG00000015501:prot
12 8 0.0032948928 >10:ENSMUSG00000019810:prot
13 1 9.852217E-4 >10:ENSMUSG00000001435:prote
14 1 7.423905E-4 >10:ENSMUSG00000005683:prote
15 2 0.0012180269 >10:ENSMUSG00000015501:prot
16 0 0.0 >10:ENSMUSG00000019810:protein_codin
17 1 9.852217E-4 >10:ENSMUSG00000001435:prote
Ln:1 Col:1 Sel:0 (0 bytes) in 0 ra Dos\Windows ANSI INS

ΚΕΦΑΛΑΙΟ 4

Συμπεράσματα

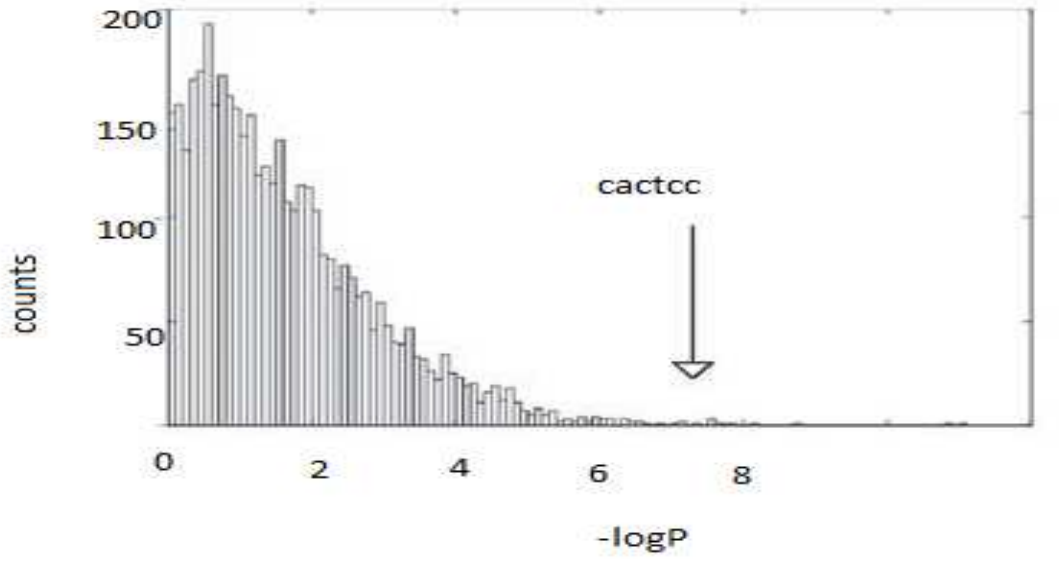
4.1 Σύγκριση Αποτελεσμάτων

Για την εξέταση των επιδράσεων που έχουν τα miRNAs στην ρύθμιση της γονιδιακής έκφρασης απαιτείται η μελέτη της αποτελεσματικής μη λειτουργικότητας και των στρατηγικών υπερέκφρασης στους ζωντανούς οργανισμούς.

Οι Jan Krutzfeldt, Nikolaus Rajewsky, Ravi Braich, Kallanthottathil G. Rajeev, Thomas Tuschl, Muthiah Manoharan & Markus Stoffel, ανέπτυξαν μια έρευνα σχετικά με την αποσιώπηση των miRNAs σε ζωντανούς οργανισμούς με την χρήση των antagomirs.[17]

Οι παραπάνω ερευνητές έχουν ανακαλύψει μια νέα κατηγορία χημικά μηχανικών ολιγονουκλεοτιδίων, τα οποία ονομάζονται «antagomirs» ως ειδικοί και αποτελεσματικοί αποσιωπητές της έκφρασης του miRNA στα ποντίκια. Αυτά τα συζευγμένα με χοληστερόλη μονόκλωνα μόρια RNA είναι νουκλεοτίδια μήκους 21-23 και συμπληρωματικά στα ολοκληρωμένα miRNAs-στόχους. Πιο συγκεκριμένα έδειξαν σιωπηλά την έκφραση miRNA (miR-122), στο ήπαρ, την καρδιά, το έντερο, το δέρμα, και τον μυελό των οστών για περισσότερο από μια εβδομάδα μετά τη χορήγηση ενδοφλέβιας ένεσης. Αυτό είχε ως αποτέλεσμα η ρύθμιση της έκφρασης γονιδίων που είχαν προβλεφθεί , να κατασταλεί από το miR-122 ,επειδή αυτά τα γονίδια είχαν ένα μοτίβο αναγνώρισης miR-122 στη με μεταφρασμένη περιοχή 3'.Παραδόξως , η θεραπεία με antagomirs αποκάλυψε επίσης έναν σημαντικό αριθμό από υπο-εκφραζόμενα γονίδια τα οποία μπορούν να ενεργοποιηθούν από το miR-122(σε αντίθεση με την καταστολή). Αν και ο μηχανισμός με τον οποίο τα miRNAs μπορούν να ενεργοποιήσουν την γονιδιακή έκφραση στους ζωντανούς οργανισμούς είναι άγνωστος ,μπορεί να συνεπάγεται μια πιο έμμεση επίδραση, και συγκεκριμένα, την καταστολή ενός μεταγραφικού καταστολέα.

Για να επιβεβαιώσουν πειραματικά τη σύνδεση μεταξύ της καταστολής και της παρουσίας των ταιριασμάτων του miR-122 του πυρήνα μέσα στην 3'UTR ,κλωνοποίησαν τις 3' UTR των πέντε γονιδίων ,τα οποία υποεκφράστηκαν απο το antagomir-122 και περιέχει μια ακολουθία miR-122 του πυρήνα σε ένα σύστημα αναφοράς λουσιφεράσης.Από τις 108 μεταγραφές που ήταν σημαντικά υποεκφραζόμενες ,παρατηρήθηκε ότι η πιθανότητα της υπόκρυψης του miR-122 πυρήνα μειώθηκε σχεδόν κατά 2.7 φορές.Για την περαιτέρω ανάλυση ,για το αν δηλαδή η υπερέκφραση ή η υποέκφραση των ακολουθιών του miR-122 πυρήνα είναι συγκεκριμένες , ανέλυσαν την παρουσία όλων των πιθανών εξαμερών μοτίβων στις υποεκφραζόμενες, υπερεκφραζόμενες και μη αλλαγμένες μεταγραφές.Όταν έγινε η σύγκριση μεταξύ των υπερεκφραζόμενων και μη μεταβαλλόμενων γονιδίων , η ακολουθία του πυρήνα του miR-122 είναι το περισσότερο σημαντικά υπερ-αναπαριστάμενο εξαμερές.Συγκεκριμένα, ο πυρήνας του miR-122 ήταν ανάμεσα στο 1% της κορυφής των υπο-αναπαριστάμενων μοτίβων για τις υποεκφραζόμενες μεταγραφές.Τα αποτελέσματα αυτά δείχνουν ότι τα υποεκφραζόμενα mRNAs στοχοποιούνται και καταστέλλονται άμεσα από το miR-122, αλλά επίσης , ότι ένας σημαντικός αριθμός υποεκφραζόμενων γονιδίων μπορούν να ενεργοποιούνται από το miR-122.

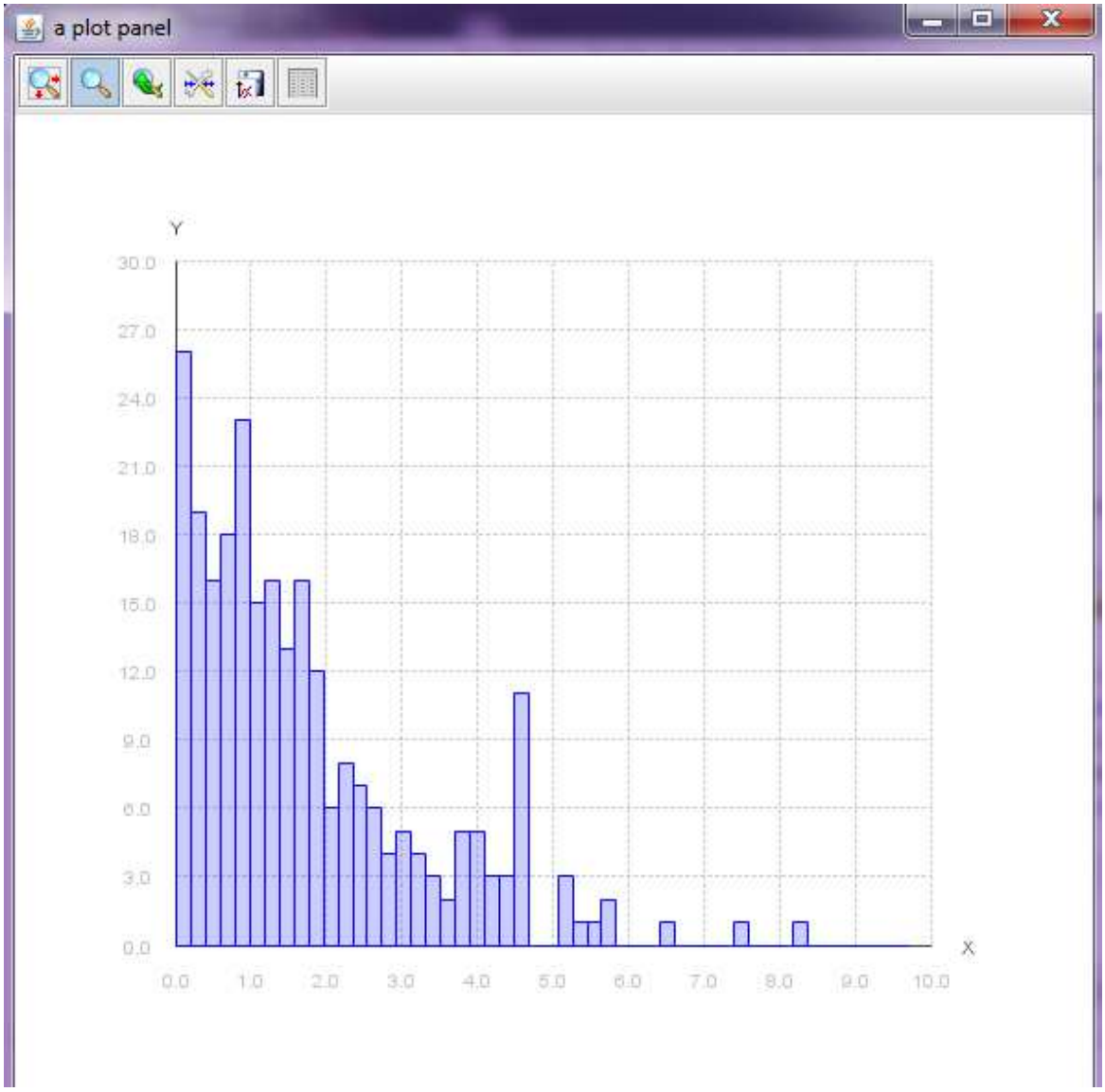


Εικόνα

Περιγραφή διαγράμματος

Για κάθε ένα από τα 4096 πιθανά εξαμερή μοτίβα RNA και κάθε μεταγραφή ,καταγράφηκε ο αριθμός των μη επικαλυπτόμενων παρουσιάσεων διαιρεμένος από το μήκος των 3'UTR.Για κάθε μοτίβο ένα μη παραμετρικό τεστ το Wiclox rank sum test εφαρμόστηκε στις κτατνομές ανάμεσα στις Upregulated και στις Unchanged μετγραφές.

Παρατηρούμε πως η κατανομή του ιστογράμματος είναι ίδια (ακολουθεί την εκθετική κατανομή) με την κατανομή του δικού μας ιστογράμματος.


a plot panel
Y
30.0
27.0
24.0
21.0
18.0
15.0
12.0
9.0
6.0
3.0
0.0
X
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0

4.1.1 Δυσκολίες και Προβλήματα

Κατά την ανάπτυξη της εφαρμογής υπήρξαν κάποια τεχνικά προβλήματα ,τα περισσότερα ευτυχώς κατάφεραν να αντιμετωπιστούν.

Αρχικά η κυριότερη δυσκολία ήταν το στατιστικό πακέτο R λόγω της μη οικειότητας με το περιβάλλον και συνέβαιναν λάθη τα οποία δεν μπορούσαν να αντιμετωπιστούν άμεσα.

Επίσης

4.2 Προτάσεις για βελτίωση και επέκταση της εφαρμογής

Όπως έχουμε ήδη αναφέρει η συγκεκριμένη εφαρμογή έχει την απαραίτητη τεχνογνωσία και έχει θέσει τα απαραίτητα θεμέλια για την υλοποίηση ενός διαδικτυακού περιβάλλοντος για ανίχνευση δυνητικών λειτουργικών περιοχών γονιδίων.

Λόγω του θέλαμε να δώσουμε περισσότερη προσοχή στην λειτουργικότητα της εφαρμογής ,που είναι κυρίως η σύγκριση των ακολουθιών και η στατιστική επεξεργασία των δεδομένων, η ενασχόληση με την εμφανίση του site έχρηζε δευτερεύουσας σημασίας.

Γι'αυτό τον λόγο μια πιθανή εξέλιξη της εφαρμογής θα μπορούσε να είναι η βελτίωση της εμφάνισης και της λειτουργικότητας της διεπαφής.

Μια άλλη πιθανή βελτίωση θα μπορούσε να είναι η σύνδεση με βάσεις δεδομένων ,από τις οποίες θα μπορεί ο χρήστης να επιλέγει τα UTRs προς σύγκριση μέσω του κωδικού τους ή του ονόματος τους και έτσι δε θα χρειάζεται να κάνει upload τα αρχεία των ακολουθιών

Θα μπορούσε επίσης να υπάρξει η δυνατότητα για οπτικοποίηση των αποτελεσμάτων,για παράδειγμα ο χρωματισμός των ν-μερών ενδιαφέροντος στις ακολουθίες και η δυνατότητα αναζήτησης «ελαστικών μοτίβων» (π.χ σε κάποια θέση του nmer να επιτρέπεται να υπάρχει κάποια διαφορετική βάση).

4.3 Επίλογος

Τα τελευταία έτη οι υπολογιστές κατακτούν σημαντική θέση σε κάθε τομέα της ζωής μας αλλά και σε αρκετούς τομείς διαφόρων επιστημών. Η Βιοπληροφορική αποτελεί ένα σύγχρονο τομέα έρευνας και ανάπτυξης τόσο για τους μοριακούς βιολόγους όσο και για τους επιστήμονες της πληροφορικής. Η συνεργασία των δύο αυτών επιστημών χαρακτηρίζεται αρκετά υποσχόμενη και με ιδιαίτερη σημασία αφού έρχεται να ρίξει φως στην ερμηνεία και το ρόλο της γονιδιακής πληροφορίας και κατ' επέκταση σε αρκετές διαδικασίες της ζωής που ζητούν ερμηνεία.

Η Βιοπληροφορική είναι αυτή η επιστήμη η οποία παρέχει τα εργαλεία και τις μεθόδους τα οποία υποστηρίζουν την ανάγκη για την εκμετάλλευση υπολογιστικής ισχύος και την εξαγωγή γνώσης από βιολογικά δεδομένα.

Η έρευνα σε αυτήν την περιοχή περιλαμβάνει την ανάλυση γενετικής/γονιδιωματικής πληροφορίας, με στόχο την πρόβλεψη, ή τον ακριβή καθορισμό βιολογικών λειτουργιών.

Με την παρούσα διπλωματική θελήσαμε να δημιουργήσουμε μια φιλική προς το χρήστη διαδικτυακή εφαρμογή για την αναζήτηση των στατιστικά υπερεκπροσωπούμενων nmers μεταξύ 2 ομάδων ακολουθιών μη μεταφραζόμενων περιοχών γονιδίων με διαφορετικό προφίλ έκφρασης, με σκοπό τον εντοπισμό δυνητικών ρυθμιστικών μηχανισμών.

.Θεωρούμε ότι η συγκεκριμένη εργασία θα μπορέσει να αποτελέσει βάση για το συγκεκριμένο αντικείμενο και θα μπορέσει να επεκταθεί κατάλληλα προσφέροντας μεγαλύτερη λειτουργικότητα.

4.4 Βιβλιογραφία

1.Pearson H (2006). "Genetics: what is a gene?". *Nature* **441** (7092): 398–401. doi:10.1038/441398a. PMID 16724031.

2.Elizabeth Pennisi (2007). "DNA Study Forces Rethink of What It Means to Be a Gene". *Science* **316** (5831): 1556–1557. doi:10.1126/science.316.5831.1556. PMID 17569836

3.Witzany G (2009). Noncoding RNAs: Persistent Viral Agents as Modular Tools for Cellular Needs. Ann NY Acad Sci. 1178:244-267.

5.Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (May 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nat. Methods* **5** (7): 621.

6. Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning, 21, 51–80.

7. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, 262, 208–214.

8. Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol., 16, 939–945.

9. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15, 563–577.

10. Saurabh Sinha and Martin Tompa YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation

11. Caselle M, Di Cunto F, Provero P: Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. BMC Bioinformatics 2002, 1:3–7.

12. Davide Cora', Ferdinando Di Cunto, Paolo Provero, Lorenzo Silengo, Michele Caselle Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs.

13. Crispin Roven and Harmen J. Bussemaker REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data

14.Martin Tompa,Nan Li,Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Anderi A Minorov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Regnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng,Christofer Workman, Chun Ye & Zhou Zhu. Assesing computational tools for the discovery of transcription factor binding sites.

15. Joerg Mattes, Ming Yang, and Paul S. Foster Regulation of MicroRNA by Antagomirs A New Class of Pharmacological Antagonists for the Specific Regulation of Gene Function?

16. . Caselle M, Di Cunto F, Provero P Identification of candidate regulatory sequences in mammalian 3' UTRs by statistical analysis of oligonucleotide distributions

17. Jan Krutzfeldt, Nikolaus Rajewsky, Ravi Braich, Kallanthottathil G. Rajeev, Thomas Tuschl, Muthiah Manoharan & Markus Stoffel, Silencing of microRNAs in vivo with 'antagomirs'

18. http://en.wikipedia.org/wiki/DNA_binding_site

19. http://en.wikipedia.org/wiki/Transcriptional_regulation

20. http://en.wikipedia.org/wiki/Gene

21. http://en.wikipedia.org/wiki/Gene_expression

22. http://en.wikipedia.org/wiki/Regulation_of_gene_expression

23. http://en.wikipedia.org/wiki/DNA

24. http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

25. http://en.wikipedia.org/wiki/P-value

26. http://www.sgenomics.org/RepAnalyse/thesis.pdf

27. http://en.wikipedia.org/wiki/Nucleic_acid_sequence

28. http://en.wikipedia.org/wiki/DNA_sequencing

29. http://www.biomedcentral.com/1471-2105/8/S7/S21

30. http://www.almob.org/content/1/1/8

31. http://forums.devshed.com/java-help-9/why-use-jsp-8590.html

32. http://java.sun.com/products/jsp/whitepaper.html

33. http://blog.revolutionanalytics.com/2009/01/using-r-as-a-scripting-language-with-rscript.html

34. http://oreilly.com/catalog/javadata/chapter/ch04.html#40793

35. http://www.servlets.com/soapbox/problems-jsp.html

36. http://genome.cshlp.org/content/12/5/739.full

37. http://www.nature.com/nbt/journal/v24/n4/full/nbt0406-423.html